Master Thesis

“Associatie analyse op profielen van sociale netwerksites”

Auteur	:	Kristel Rösken
Datum	:	30-01-2010
Studentennummer	:	0546070
Scriptienummer		116 IK
Examinator / Begeleider	:	Tom Heskes
Onderwijsinstelling	:	Radboud Universiteit Nijmegen
Opleiding	:	Informatiekunde
Afstudeerbedrijf	:	Logica Arnhem
Bedrijfsbegeleider	:	Raynni Jourdain
Bedrijfsmentor	:	Bertram Kolhoff
Opdrachtgever	:	Ivo van der Heijden

Gebaseerd op onderzoek uitgevoerd in de periode augustus 2007 t/m februari 2008

Inhoud Kort

Dit document is gebaseerd op onderzoek uitgevoerd in de periode augustus 2007 t/m februari 2008.

Overal in dit document waar ‘hij’ of ’zijn’ staat, mag ook het vrouwelijke equivalent ‘zij’ of ‘haar’ worden gelezen. Om dit document leesbaar te houden is ervoor gekozen zo veel mogelijk één variant te gebruiken.

1 Samenvatting

Sociale netwerksites worden inmiddels door ongeveer driekwart van de volwassen Nederlanders gebruikt. Ook in het buitenland zijn sociale netwerksites erg populair. De sites worden vooral gebruikt om nieuwe vriendschappen te maken en bestaande vriendschappen te onderhouden. Wanneer mensen lid worden van het netwerk maken zij een profiel aan met persoonlijke informatie. Al deze profielen samen vormen een schat aan informatie over de gebruikers, die gebruikt kunnen worden om gericht te adverteren op de sites.

Om gericht te kunnen adverteren is het belangrijk te weten of de data die men daarvoor gebruikt betrouwbaar is. Mensen hebben verschillende redenen om te liegen op hun profiel, o.a.: om de privacy te beschermen, om toegang te verkrijgen of om te voldoen aan de heersende norm. Ook kunnen gegevens verouderd zijn of per ongeluk foutief ingevuld zijn. Daarnaast zijn er mensen die meerdere profielen hebben op één sociale netwerksite.

Uit de enquête die is uitgevoerd in het kader van deze thesis is gebleken dat negen procent van de gebruikers meer dan één profiel heeft op de sociale netwerksite waarover zij de enquête hebben ontvangen. Zestien procent van de respondenten gaf aan dat van de twaalf gevraagde eigenschappen er minstens één in hun profiel niet klopte. Wanneer naar alle informatie op het profiel wordt gekeken, geeft negentien procent aan minstens één fout op zijn profiel te hebben staan. Woonplaats, opleiding en relatiestatus staan het meest verkeerd vermeld op het profiel. Sterrenbeeld en geslacht zijn vaker waarheidsgetrouw.

Om de mogelijkheden voor het gericht plaatsen van advertenties uit te breiden en mogelijk te verbeteren, is een experiment uitgevoerd waarin datamining werd toegepast op de profielen van twee sociale netwerksites. Binnen wetenschappelijk onderzoek werd al eerder datamining toegepast op sociale netwerksites, maar daarin werd gekeken naar de connecties tussen mensen. Dit onderzoek leek aan het begin van de onderzoeksperiode het eerste te zijn waarin datamining werd toegepast op de inhoud van de profielen op sociale netwerksites.

Voor het uitvoeren van datamining is gebruik gemaakt van het in CBA geimplementeerde Apriori algoritme om associatie analyse uit te voeren. Dit levert associatieve regels als resultaat. Aan de hand van zes kwaliteitsmaten en de beoordeling van vier HEAO docenten op het nut voor adverteerders, zijn de beste regels geselecteerd.

De ruim vierhonderd resulterende associatieregels hebben een supportwaarde van minimaal één procent, een confidencewaarde van minimaal 80 procent en zijn door alle docenten als nuttig of zeer nuttig aangemerkt.

Uit dit onderzoek blijkt dat d.m.v. associatie analyse regels gegenereerd kunnen worden die geschikt zijn om gericht mee te adverteren. Verder wetenschappelijk onderzoek is wenselijk om de conclusies van dit onderzoek verder te valideren en te onderzoeken in hoeverre gericht adverteren op sociale netwerksites aan de hand van associatie analyse betere resultaten oplevert.

2 Voorwoord

Ruim een half jaar voor de start van mijn onderzoek, kwam ik een artikel tegen dat in Melbourne's, The Age newspaper had gestaan. Dit artikel[1] inspireerde mij voor het onderwerp van deze thesis. In het betreffende artikel wordt beschreven dat trendspotters steeds vaker blogs gebruiken om nieuwe (kleding)trends te spotten. Zij doen dit door zelf naar blogs te surfen en deze te lezen en bekijken. Hieronder vind u een citaat van het artikel.

"On the frontlines of cool, bloggers with street cred are the unwitting quarry of forecasters anxiously scanning the net in the hope of spotting just one thing: the next big trend.The forecasters have a wide net to cast in: there are an estimated 100 million blogs worldwide, about 450,000 of which are in Australia. "

Het artikel schetst het grote aantal blogs, dat te groot is voor een trendspotter om deze eenvoudig te scannen op nieuwe trends. Een van de mogelijkheden van datamining is het ontdekken van trends in de data. Datamining zou een ondersteunende rol kunnen vormen in het spotten van trends op blogs. Het geautomatiseerd spotten van modetrends spotten op foto's is behoorlijk complex, maar het vinden van trends in tekst dmv datamining is gemeengoed. Naast trends op blogs zijn er ook andere plekken waar mensen hun mening geven (bijvoorbeeld sociale netwerksites) waarvan de data ook als bron gebruikt kan worden.

De reden voor het zoeken naar modetrends is duidelijk, maar wat kan er gedaan worden met trends in onderwerpen op blogs of sociale netwerken? Het gebruik van de resultaten van datamining kan o.a. gebruikt worden voor het plaatsen van advertenties. Wanneer blijkt dat een bepaald product veel genoemd wordt onder mensen met een bepaalde leeftijd, geslacht en hobby, kan er voor gekozen worden bij alle mensen met die leeftijd, geslacht en hobby een advertentie voor een aanbieder van dat product te plaatsen, ook wanneer zij niet dit product noemen op hun profiel. Ten tijde van dit onderzoek werden leeftijd, geslacht en woonplaats al gebruikt bij gericht adverteren op sociale netwerksites, maar overige informatie nog niet.

Informatie dient correct te zijn om succesvol gericht te adverteren. Voor specifieke doeleinden (bijvoorbeeld advertenties voor virituele producten[2]) kan het interessanter zijn hoe een persoon zich online voordoet. Echter voor de meeste adverteerders is het belangrijker de correcte eigenschappen van gebruikers in de fysieke wereld te kennen. Daarom zal door middel van een databetrouwbaarheidsonderzoek worden getracht de betrouwbaarheid van de gegevens van gebruikers te achterhalen. Deze onderzoeksresultaten kunnen gebruikt worden bij de evaluatie van een experiment waarin datamining wordt toegepast op profielen van sociale netwerksites. Het doel is betrouwbare associatieve regels op te leveren die gebruikt kunnen worden om gericht te adverteren.

[1] = http://www.theage.com.au/news/web/coolhunting-on-the-web/2006/09/20/1158431719559.html?page=fullpage#contentSwap1
[2] = Onder andere gebruiksvoorwerpen voor de avatar in een spel als Second Life.

3 Inleiding

Sociale netwerksites werden in 2007 door ongeveer driekwart van de volwassen Nederlanders gebruikt. Ook in het buitenland zijn veel mensen lid van sociale netwerksites. Wanneer mensen lid worden van een dergelijke site maken zij een profiel aan met persoonlijke informatie. Al deze profielen samen vormen een schat aan informatie over de gebruikers, deze kan gebruikt worden om gericht te adverteren op de sites.

Dit document is gebaseerd op onderzoek uitgevoerd in de periode augustus 2007 t/m februari 2008. Binnen dit onderzoek is de betrouwbaarheid van gegevens op de profielen bekeken. Daarnaast wordt een experiment beschreven dat uitgevoerd is om te testen of het toepassen van datamining een verbetering of toevoeging kan vormen op de huidige manieren om gericht te adverteren.

Het doel van dit onderzoek is te bekijken of het gebruik van associatieve regels gegenereerd door middel van datamining, mogelijk een bruikbare aanvulling zou kunnen zijn op het gericht adverteren op sociale netwerksites. De commerciële waarde van deze aanvulling valt buiten de scope van dit onderzoek.

Allereerst wordt in hoofdstuk 4 een korte inleiding beschreven wat er in dit onderzoek aan bod komt en op welke onderwerpen gerelateerd wetenschappelijk onderzoek heeft plaatsgevonden.

Vervolgens wordt een algemene introductie gegeven in de wereld van de sociale netwerksites. In hoofdstuk 5 wordt besproken wat sociale netwerksites zijn, wat de content van de sites is, wie deze sites gebruiken en hoe de sites gebruikt worden.

Door middel van een literatuuronderzoek, zijn aanwijzingen gevonden over de betrouwbaarheid van sociale netwerken, deze zijn gedocumenteerd in hoofdstuk 6. De betrouwbaarheid is vervolgens onderzocht in de vorm van een enquête onder gebruikers van vier sociale netwerksites. De aanpak van dit betrouwbaarheidsonderzoek is beschreven in hoofdstuk 7 en de resultaten zijn terug te vinden in hoofdstuk 8.

Een experiment is uitgevoerd om te onderzoeken hoe door middel van datamining associatieve regels gegenereerd kunnen worden met als doel deze te gebruiken voor het gericht adverteren op sociale netwerksites. In hoofdstuk 9 is uiteengezet in welke vorm datamining wordt gebruikt op bronnen gerelateerd aan het internet en welke vormen al in eerder wetenschappelijk onderzoek gebruikt zijn op sociale netwerksites. Ook is in hoofdstuk 9 de keuze voor het algoritme uiteengezet. In hoofdstuk 10 zijn de aanpak en de voorbereidingen van het experiment beschreven. In hoofdstuk 11 worden de resultaten van het experiment naar voren gebracht.

Hoofdstuk 12 combineert de belangrijkste bevindingen uit het literatuuronderzoek, het betrouwbaarheidsonderzoek en het datamining-experiment en vat deze samen. Hoofdstuk 13 vormt een discussie over de aanpak en de uitkomsten van dit onderzoek. Aanbevelingen voor verder wetenschappelijk onderzoek worden beschreven in hoofdstuk 14. Hoofdstuk 15 vormt de conclusie van dit onderzoek.

De personen die een bijdrage geleverd hebben ter ondersteuning van dit onderzoek bedank ik in hoofdstuk 16. Alle in deze thesis genoemde sociale netwerksites worden opgesomd in hoofdstuk 17. In hoofdstuk 15 wordt de betekenis gegeven van de in dit document gebruikte begrippen die mogelijk niet bij iedere lezer bekend zijn. De bronnen waarnaar in dit document gerefereerd is, worden in hoofdstuk 16 opgesomd. In deze thesis wordt naar artikelen over onderzoeksresultaten gerefereerd met de letter ‘O’, referenties naar teksten uit online media, offline media, persberichten e.d. zijn weergegeven met de letter ‘M’.

4 Dit onderzoek

In de volgende hoofdstukken zal onderzoek worden gedaan in de praktijk. Dit onderzoek focust op databetrouwbaarheid en datamining van profielen. Deze focus is gelegd om aan de hand van de onderzoeksresultaten achter te komen of associatie analyse een goede methode is om gericht te adverteren op sociale netwerksites. Om gericht te kunnen adverteren is het belangrijk ook te weten of de data die men daarvoor gebruikt betrouwbaar is.

Buiten de scope van dit onderzoek vallen:

- Validatie van de vragenlijst voor het databetrouwbaarheidsonderzoek.
- Correctie van onderzoeksresultaten naar afspiegeling van de sociale netwerksite.
- Correctie van onderzoeksresultaten naar afspiegeling van de samenleving.
- Analyse van verschillen in databetrouwbaarheid tussen verschillende sociale netwerksites.
- Onderzoek naar eigenschappen van sites welke invloed hebben op databetrouwbaarheid.
- Het onderzoeken van de commerciële waarde van inzet van de beschreven datamining technieken.
- Het gebruiken van de semantische oriëntatie van woorden binnen het datamining experiment.
- Het uittesten van verschillende datamining algoritmen op de data, het datamining experiment wordt uitgevoerd voor één algoritme.
- Onderzoek naar kwaliteitsmaten voor het meten van het resultaat van datamining. Kwaliteitsmaten zijn gebruikt, maar niet diepgaand onderzocht.

4.1 Opbouw onderzoek

Deze thesis bevat twee praktijk onderzoeken, namelijk een enquête over de databetrouwbaarheid en een experiment met associatie analyse op data uit profielen. De beschrijving van aanpak en resultaten van de praktijkonderzoeken wordt voorafgegaan door een theoretische oriëntatie in de vakgebieden die in de praktijkonderzoeken aan bod komen, o.a. sociale netwerksites en associatie analyse. Ten slotte worden de resultaten van de onderzoeken geanalyseerd.

4.1.1 Sociale netwerksites

Als basis voor de praktijkonderzoeken wordt eerst gekeken naar de definitie van sociale netwerksites, de content op de sites, de gebruikers en de mogelijkheden met betrekking tot adverteren aan de hand van literatuur.

4.1.2 Databetrouwbaarheid

Binnen het databetrouwbaarheidsonderzoek wordt eerst bronnen uit de literatuur aangeboord om een beeld te krijgen van de redenen van leugens en fouten op sociale netwerksites en de mate waarin deze voorkomen. Vervolgens worden leden van een aantal sociale netwerksite benaderd via email voor een enquête. In deze enquête wordt hun gevraagd welke foutieve data zij op hun profiel hebben staan.

4.1.3 Associatie analyse

Om een basis te leggen voor het associatie analyse experiment, wordt bekeken wat associatie analyse is, hoe een associatie analyse algoritme werkt en welke andere methoden van datamining er zijn. Ook worden onderzoeken genoemd waarin datamining op online content en/of persoonlijke eigenschappen is toegepast.

Hierna wordt de data van profielen van twee sociale netwerksites geschikt gemaakt voor associatie analyse. Het Apriori algoritme wordt toegepast, waarvan de geschikte resultaten door middel van kwaliteitsmaten en de commerciële potentie worden geselecteerd.

4.1.4 Analyse resultaten

Ten slotte worden resultaten van beide onderzoeken geëvalueerd en gecombineerd. Aanbevelingen voor het gebruik van associatie analyse op profielen van sociale netwerksites en toekomstig wetenschappelijk onderzoek op dit gebied worden beschreven.

4.2 Eerder wetenschappelijk onderzoek

De onderzoeksgebieden waar deze thesis zich op focust zijn nog weinig aan bod gekomen in eerder wetenschappelijk onderzoek. Om dit onderzoek toch te kunnen relateren aan bevindingen van anderen, is ook gebruik gemaakt van niet wetenschappelijke bronnen zoals krantenartikelen.

4.2.1 Sociale netwerksites

Een groot deel van het wetenschappelijk onderzoek naar sociale netwerksites (uitgebracht voor 2008) focust op een van de volgende onderwerpen [O1]:

- Impressie management
- De betekenis van 'vriendschap' op een sociale netwerksite
- Netwerken en netwerkstructuur
- De brug tussen online en offline connecties
- Privacy

Niet al deze onderwerpen zijn van belang voor dit onderzoek. Artikelen over sociale netwerksites die relevant zijn voor dit onderzoek komen aan bod in hoofdstuk 5 en 6.

4.2.2 Databetrouwbaarheid

Databetrouwbaarheid van profielen is al in verschillende onderzoeken [O2], [O3], [O4], [O5] en [O6] aan bod gekomen. In deze onderzoeken wordt veelal niet direct gesproken over de databetrouwbaarheid, vaak gaat het om redenen voor leugens en identiteitsexperimenten op sociale netwerksites.

4.2.3 Associatie analyse

De meeste wetenschappelijke onderzoeken naar associatie analyse focussen zich op een van de volgende onderwerpen:

- Introductie van algoritmes
- Verbetering van bestaande algoritmes
- (nieuwe) Toepassingen van associatie analyse
- Combinatie van datamining/associatie analyse technieken en technieken uit aangrenzende vakgebieden (bijvoorbeeld Information Retrieval)

Dit onderzoek kunnen we in de derde categorie onderbrengen, omdat dit onderzoek gaat om het toepassen van associatie analyse op sociale netwerksites.

Wetenschappelijk onderzoek naar datamining en associatieanalyse dat voor dit onderzoek relevant is, komt aan bod in hoofdstuk 9, 10 en 11.

5 Sociale netwerksites

Dit hoofdstuk dient als inleiding en fundering van de volgende hoofdstukken. We leggen een basis ter begripsvorming voor de praktijkonderzoeken. Dit hoofdstuk geeft een indruk van het doel van sociale netwerksites, de reden voor de groei, de inhoud op deze sites, de mensen die deze sites gebruiken en hoe zij de sites gebruiken.

5.1 Definitie

5.1.1 Sociale netwerksites

Een sociale netwerksite soms ook wel profielensite [M1], profielsite [M2], online smoelenboek [M1] of digitaal vriendennetwerk [M3] genoemd. De term sociale netwerksite wordt ook wel eens afgekort tot SNS [O7]. Dezelfde afkorting wordt ook wel gebruikt voor andere begrippen, waaronder sociale netwerk services [O8].

Boyd en Ellison [O1] definiëren sociale netwerksites als een web-based service die het individuen mogelijk maakt:

- een publiek of semi-publiek profiel aan te maken binnen een gebonden systeem
- een lijst samen te stellen van anderen waarmee zij een connectie hebben
- de lijst met connecties van henzelf en van anderen in het systeem te bekijken en te bezoeken. De benaming en karakteristieken van deze connecties kunnen per site verschillen.

Vos [O9] definieert sociale netwerksites als virtuele omgevingen waarin mensen een profiel creëren waarin ze zichzelf beschrijven en zichzelf via de profielen verbinden met andere mensen op de site die ze kennen, waardoor ze een netwerk van persoonlijke connecties creëren. Dit netwerk van persoonlijke connecties is op deze sites zichtbaar en vormt het belangrijkste element van de wijze waarop mensen zichzelf presenteren.

Beide definities hebben grote overeenkomstigheden en sluiten elkaar niet uit. De definitie van Boyd en Ellison is het meest volledig en duidelijk. Daarom is gekozen de definitie van Boyd en Ellison te gebruiken voor dit onderzoek. Hierbij wordt wel de kanttekening gemaakt dat sommige sociale netwerksites de acties uit de opsomming van Boyd en Ellison niet alleen mogelijk maken voor individuen maar ook voor groepen, organisaties of merken. Echter sites die één of meerdere van deze acties niet mogelijk maken voor individuen, vallen buiten de in dit onderzoek gebruikte definitie.

Hoewel er veel sites bestaan die functionaliteiten gemeen hebben met sociale netwerksites, differentiëren sociale netwerksites zich door de manier waarop er met connecties, profielen en commentaar/berichten om gegaan wordt van andere sites. Het grootste verschil met andere sites is de publieke zichtbaarheid van de communicatie tussen connecties [O3]. Vaak kan men op sociale netwerksites ook communiceren zonder dat dit door derden te zien is. De publieke zichtbaarheid van connecties van een persoon en de veelgebruikte publiekelijk te communiceren, differentiëren sociale netwerksites van andere sites.

De grootste sociale netwerksites maken geen keuze voor een specifieke doelgroep, of dragen deze niet uit. Sites die wel een keuze voor een specifieke doelgroep hebben gemaakt en uitdragen, definiëren de doelgroep bijvoorbeeld aan de hand van: leeftijd, geslacht, ras, seksuele voorkeur, beroepsgroep, geloofsovertuiging, (sociale) klasse of hobby's.

5.1.2 Profiel

Een profiel op een sociale netwerksite is een zelfgeschreven beschrijving van een individu. Het is gebruikelijk dat een dergelijke beschrijving onder andere naam, e-mailadres, geslacht, woonplaats, geboortedatum, foto's, connecties/vrienden en interesses bevat. Deze onderwerpen worden niet altijd bij iedereen belicht en er zijn nog tal van andere onderwerpen die ook vaak op een profiel aan bod komen. Een uitgebreidere lijst staat in de paragraaf 'Content'.

Volgens Boyd [O3] is de stijl van de sociale netwerksites afkomstig van datingsites. Hierdoor bevat een profiel vaak materiaal dat typisch bij datingsites hoort zoals demografische details (leeftijd, geslacht, locatie, e.d.), smaak (interesses, hobby's favoriete bands e.d.) en een foto. Op datingsites worden wel profielen gebruikt, maar een verschil is dat mensen vaak niet hun echte naam gebruiken [O10], connecties tussen mensen niet zichtbaar zijn [O10] en het niet gebruikelijk is publiek te communiceren.

Zoals al in de definitie van een profiel aangegeven is, kan de gebruiker bij het aanmaken van een profiel, beslissen of dit publiek of semi-publiek profiel is. Een profiel is bij de meeste sociale netwerksites standaard publiek, dat wil zeggen dat het voor iedereen (vaak alleen de leden van de site) toegankelijk is, zolang de gebruiker dit niet beperkt heeft. Veertig procent van de Amerikaanse online jongeren heeft een profiel dat iedereen kan zien [O11]. Een groot deel van de sociale netwerksites biedt gebruikers de mogelijkheid om het gehele profiel of een bepaald deel van de gegevens op het profiel, alleen voor een bepaalde groep mensen zichtbaar te laten zijn [O3]. Dit wordt een semi-publiek profiel genoemd. Van de Amerikaanse online jongeren heeft 59 procent een semi-publiek profiel, dat bijvoorbeeld alleen zichtbaar is voor vrienden [O11].

Na het aanmaken van een profiel wordt gebruikers gevraagd hun vrienden/connecties uit te nodigen, dit kan meestal door hun e-mailadressen in te vullen of door een knop op de profielen van hun vrienden op de site aan te klikken. Bij de meeste sociale netwerksites is goedkeuring van twee personen nodig voordat deze connectie op de site staat [O3].

5.2 Content

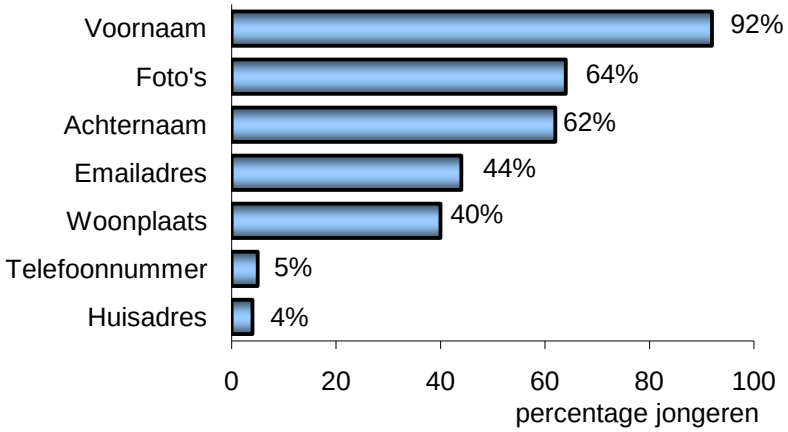
Een profiel van een individu kan o.a. de volgende persoonlijke data bevatten [O12]:

- Naam
- Geboortedatum
- Geslacht
- E-mail adres
- Woonadres
- Postcode
- Woonplaats
- Land
- Telefoonnummer
- Foto
- Onderwerp opleiding
- Opleidingsniveau
- Naam van onderwijsinstituut
- Website
- Seksuele voorkeur
- Relatiestatus
- Interesses
- Beroep
- Werkgever
- Favoriete muziek
- Favoriete boeken
- Favoriete films
- Favoriete TV-series
- Persoonlijk motto
- Netwerk van vrienden/connecties
- Deelname aan groepen

5.2.1 Meest ingevoerde content

Niet alle hierboven genoemde gegevens worden door alle gebruikers op hun profiel ingevuld. Uit figuur 1 is af te lezen hoeveel procent van de Nederlandse jongeren tussen 12 en 18 jaar bepaalde gegevens online heeft staan. Deze cijfers zijn afkomstig uit onderzoek van Digibewust en Stichting Mijn Kind Online [M3]. Deze percentages zijn mogelijk lager voor de gegevens die zij op sociale netwerksites hebben staan.

Johnes en Soltren [O13] geven aan dat demografische data gecombineerd met interesses, waarschijnlijk het meest interessant is voor adverteerders.



Figuur 1: Gegevens die nederlandse jongeren online hebben staan [M3]

5.2.2 Zichtbaarheid van informatie

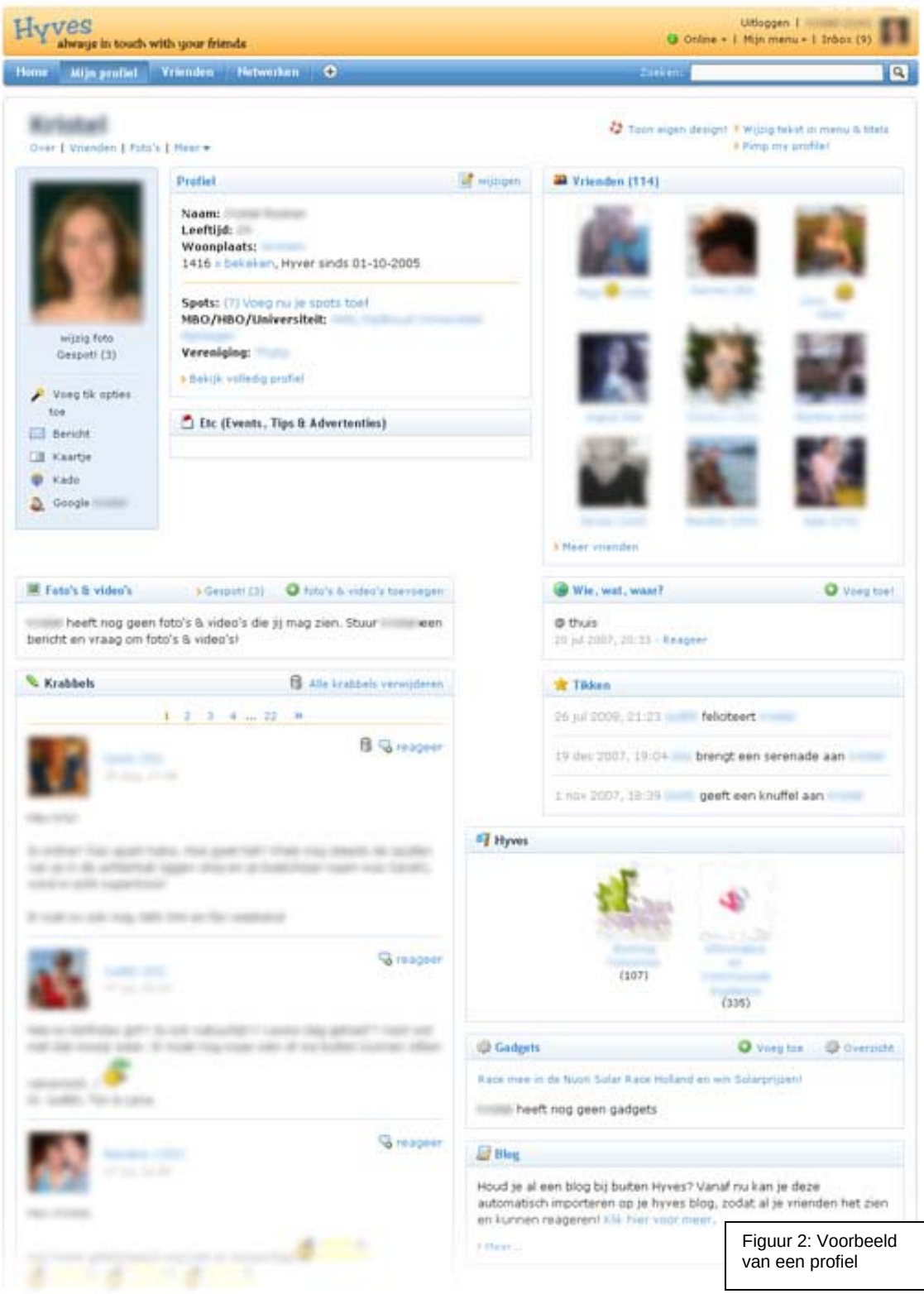
Nederlandse meisjes zeggen vaker dan jongens bepaalde gegevens af te schermen en niet alles over zichzelf te vertellen [M3]. Amerikaanse studentes blijken hun contactgegevens minder te publiceren dan mannelijke studenten, maar vertellen vaker iets over hun interesses [O13].

5.2.3 Verplichte content

Het invoeren van een aantal gegevens wordt meestal verplicht gesteld. Vaak wordt het invoeren van deze gegevens verplicht gesteld omdat de beheerder van de site ze verzameld voor marketingonderzoek en het trekken van adverteerders. Soms zitten hier echter andere redenen achter, bijvoorbeeld (zoals bij leeftijd) wetgeving. De privacywetten op dit gebied verschillen per land. Bedrijven in de VS moeten kinderen onder de dertien jaar weren wanneer zij geen toestemming hebben van ouders [O2]. In Nederland zijn er restricties op het publiceren van gegevens van kinderen jonger dan zestien [M4]. Een aantal gevolgen van deze leeftijdsrestricties zijn weergegeven in paragraaf 6.3.3.

5.2.4 Informatie anders dan persoonsgegevens

Naast de ingevulde persoonsgegevens bevat een profiel connecties en commentaar/berichten. De weergave van connecties omvat meestal foto’s en namen/nicknames die linken naar het profiel van de connectie. Op die manier kunnen bezoekers het netwerk bekijken, door te surfen van connectie naar connectie [O3]. De connectie die men op een profiel heeft staan, bevatten niet alleen goede vrienden, maar vaak ook kennissen [O5].

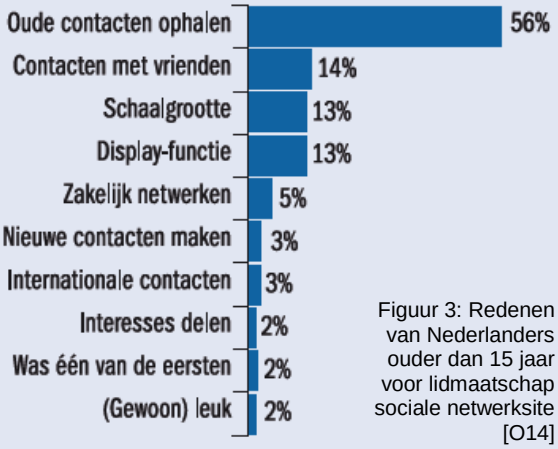


Figuur 2: Voorbeeld van een profiel

5.3 Gebruik

Sociale netwerksites worden vooral gebruikt voor:

- Contacten opdoen en onderhouden:
  Vriendschappen [O3] [O11] [O14] [O6]; relaties [O6]; zakelijke contacten [O15]
- Flirten [O11] en relaties [O6]
- Entertainment [O3]:
  Interesses delen [O14]
- Inzicht verkrijgen in een (doel)groep:
  Bijvoorbeeld door bedrijven [O5], onderzoekers of jongeren [O3]
- In contact komen met de doelgroep:
  Bijvoorbeeld door: gemeenten, politieke partijen [M5], adverteerders; drugsdealers [O5]; pornosterren [O5]; pedofielen [M6] [M7]
- Zichzelf presenteren:
  Zowel professioneel [O15] als privé [O5]
- Informatie over een individu zoeken:
  Bijvoorbeeld door: Headhunters [O5]; politie [M8]; belastingdienst [M9]; criminelen [M10], phishers [O16] [M11].
- Bepaalde content zoeken:
  Bijvoorbeeld: muziek [O17] of videomateriaal.
- Onderlinge communicatie met een specifiek doel:
  Bijvoorbeeld: elektronische leeromgeving [M12], condoleanceregister [O5] [M13]



Figuur 3: Redenen van Nederlanders ouder dan 15 jaar voor lidmaatschap sociale netwerksite [O14]

De meest voorkomende redenen van gebruik worden hieronder toegelicht.

5.3.1 contacten opdoen en onderhouden

De sites worden onder andere gebruikt om vriendschappen te onderhouden [O3], dit geld voor negentig procent van de Amerikaanse jongeren [O11]. Andere onderzoeken [O14] [O18] geven ook aan dat contact onderhouden met vrienden en het ophalen van oude contacten de activiteiten zijn waarvoor sociale netwerksites het meest gebruikt worden. Dit geld ook voor de overige leeftijdscategorieën.

Sociale netwerksites worden door bijna de helft van de amerikaanse jongeren gebruikt om nieuwe vrienden te vinden [O11]. Hoewel slechts drie procent van de Nederlandse volwassenen dit doel heeft[O14], wijst onderzoek onder Nederlandse jongeren [O6] uit dat vijfendertig procent door middel van de sites nieuwe vrienden heeft gekregen.

5.3.2 Entertainment & inzicht verkrijgen in een groep

Sociale netwerksites worden vaak gebruikt als entertainment: “Sociaal voyeurisme verdrijft verveling en geeft ondertussen inzicht in de maatschappij” [O3]. Men gebruikt deze sites om meer over hun contacten of mensen in het algemeen te weten te komen. Voorbeelden zijn bijvoorbeeld het uitzoeken wat voor een leven oud-klasgenoten nu hebben (bijvoorbeeld aan de hand van opleiding, hobby’s en relatiestatus) of het opzoeken van het meisje waar een vriend een afspraakje mee heeft.

5.3.3 Flirten & Relaties

Zeventien procent van de Amerikaanse jongeren gebruikt deze sites om te flirten [O11]. Acht procent van de Nederlandse jongeren zegt door middel van de sites een relatie te hebben gekregen [O6]. Zowel voor het flirten als voor vriendschappen geldt dat online gesprekken offline voortgezet worden en vice versa [O5].

5.3.4 Zichzelf presenteren

Profielen zijn communicatie op zich zelf. Gebruikers vertellen de massa over zichzelf. Berichten op profielen zijn gerichte communicatie. Profielen worden ook gebruikt als gespreksstof; informatie of foto's op een profiel kunnen de aanleiding vormen voor een gesprek [O5]. Het aantal reacties dat gebruikers van CU2 (een Nederlandse sociale netwerksite) ontvangen hadden op hun profiel was uiteenlopend van nul tot 350, met een gemiddelde van 25 reacties. Deze reacties waren soms negatief, maar het merendeel van de reacties was positief [O6].

5.4 Gebruikers

In deze paragraaf wordt toegelicht wat de grootste gebruikersgroepen zijn van sociale netwerksites en hoe vaak zij deze sites gebruiken.

5.4.1 Populariteit

In de loop van 2005 werden sociale netwerksites als MySpace en Facebook veelgebruikt door jongeren in de Verenigde Staten. Sociale netwerksites zijn een belangrijk onderdeel in het sociale leven van jongeren [O3]. Onderzoek uitgevoerd in 2006 [O3] wees uit dat veel Amerikaanse jongeren meenden dat aanwezigheid op deze sites essentieel was om er bij te horen op school. Ook buiten de Verenigde Staten werden sociale netwerksites al snel populair.

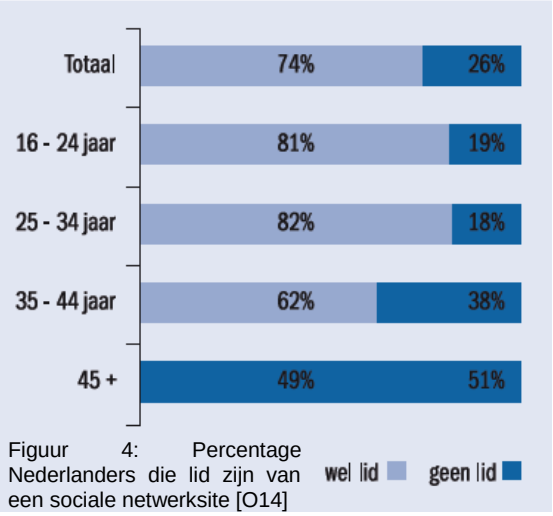
In 2007 was 74 procent van de Nederlanders ouder dan vijftien lid van een sociale netwerksite [O14]. In 2008 groeide het aantal gebruikers nog steeds, wereldwijd was de groei 5,4 procent.
Achtenveertig procent van de online Amerikaanse jeugd bezoekt sociale netwerksites gemiddeld één keer per dag of vaker. Zevenentwintig procent van de Nederlanders (vanaf vijftien jaar) die lid zijn van een sociaal netwerk, logt gemiddeld één keer per dag in [O14]. Eén op de elf online besteedde minuten wordt besteed aan sociale netwerksites [O19].

Gemiddeld zijn Nederlanders ouder dan vijftien jaar lid van 1,8 sites, waarbij ongeveer de helft lid is van één site en de andere helft van twee of meer sites [O14].

5.4.2 Leeftijdsgroepen

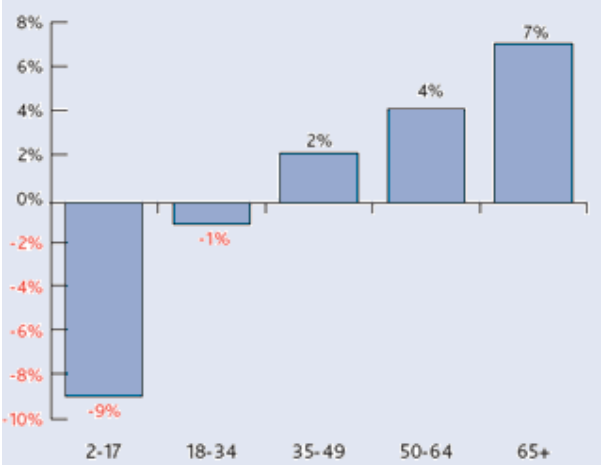
Uit onderzoek onder Amerikaanse jongeren [O11] blijkt dat oudere jongeren (leeftijd 15-17) vaker profielen aanmaken dan de jongsten (leeftijd 12-14). Het geslacht blijkt ook een rol te spelen. Jonge jongens hebben vaker een profiel dan jonge meiden (46 procent vs. 44 procent) terwijl oudere meiden vaker een profiel hebben dan oudere jongens (70 procent vs. 57 procent) [O11].

De jongeren die niet lid zijn van een sociale netwerksite willen of kunnen dit niet. Redenen hiervoor zijn: gebrek aan toegang tot internet, verbod van de ouders, principiële redenen of omdat zij de sites niet bij zichzelf vinden passen [O3].



Figuur 4: Percentage Nederlanders die lid zijn van een sociale netwerksite [O14]

Ook oudere mensen zijn vertegenwoordigd op dit soort sites. Dit gebeurt onder andere omdat jongeren hun docent of sportcoach uitnodigen ook deel te nemen aan het netwerk. Naar het verstrijken van de jaren blijft een deel van de gebruikers van het eerste uur actief, wat de totale populatie ook ouder maakt. Naar verloop van tijd zijn er veel sociale netwerksites opgericht waar men zich specifiek richt op een andere leeftijdscategorie. Voorbeelden zijn: LinkedIn.com, Vrouwzijn.nl, Boomertown.com en kinderlines.nl. Dat het publiek dat gebruik maakt van de sites wordt steeds ouder wordt, is ook te zien in figuur 5.

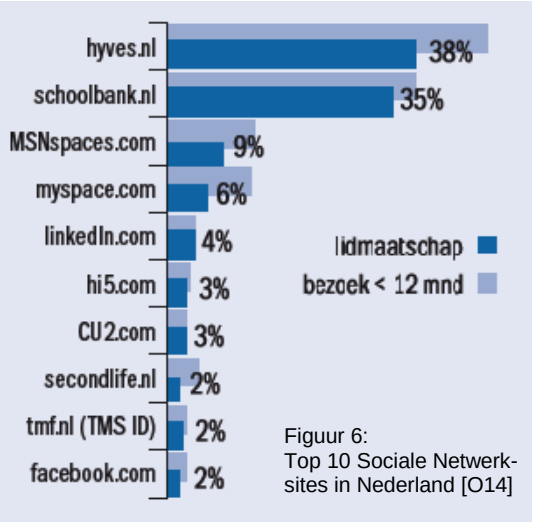


Figuur 5: Verschuiving van jong naar oud [O19]

5.4.3 Nationaliteiten en wereldwijd gebruik

Een sociale netwerksite heeft meestal gebruikers van één nationaliteit. Taalverschillen zijn hier onder andere een oorzaak van. Daarnaast overschrijdt het fysieke sociale netwerk van veel jongeren maar weinig landsgrenzen. Er zijn wel een aantal sites die succesvol meerdere nationaliteiten trekken, bijvoorbeeld Orkut, Cyworld en MySpace [O3].

Ook in Nederland zijn de meeste mensen lid van een Nederlandse sociale netwerksite. Hyves, schoolbank.nl en tmf.nl zijn van Nederlandse origine en omvatten gezamenlijk 76 procent van de Nederlanders met een lidmaatschap van een sociale netwerksite [O14]. De percentages per site zijn af te lezen in figuur 6.



Figuur 6:
Top 10 Sociale Netwerk-sites in Nederland [O14]

5.5 Adverteren

Omdat sociale netwerksites een redelijk nieuw verschijnsel zijn, wordt er veel geëxperimenteerd met de verschillende mogelijkheden van adverteren op deze sites. Tegelijkertijd zijn adverteerders enigszins terughoudend om op de sites te adverteren. Een reden hiervoor is dat zij minder controle kunnen uitoefenen op de inhoud van de pagina. Ook zijn de resultaten van adverteren op sociale netwerksites minder voorspelbaar in vergelijking met niet user generated content sites en andere media. De resultaten zijn minder voorspelbaar omdat deze manier van adverteren nog niet zo lang wordt toegepast als bijvoorbeeld het adverteren in een krant. Tegelijkertijd zijn er bedrijven die adverteren op sociale netwerken om hun doelgroep te kunnen bereiken zonder te hoeven voldoen aan de steeds strengere regelgeving met betrekking tot reclame op tv [M14].

De volgende paragrafen verschaffen meer inzichten in de inkomsten die sociale netwerksites genereren door middel van advertenties, de obstakels bij het adverteren op deze sites en de verschillende mogelijkheden. Dit wordt uiteengezet omdat het doel van dit onderzoek is door middel van associatie analyse de advertentiemogelijkheden te vergroten en te verbeteren.

5.5.1 Inkomsten

Uit onderzoek van Williamson [M15] blijkt dat sociale netwerksites wereldwijd gezamenlijk in 2007 ongeveer 1,2 miljard dollar ontvingen aan inkomsten uit advertenties. In 2008 is dit gegroeid naar twee miljard. Voor 2009 wordt een bedrag van 2,35 miljard verwacht [M16]. Williamson denkt dat de advertentie inkomsten blijven stijgen. Wel heeft ze haar verwachting gecorrigeerd als gevolg van de neergaande economie. In 2007 [M15] schatte zij namelijk hogere bedragen in dan in 2009 [M16]. Bij het vergelijken van winstcijfers moet er rekening mee gehouden dat het aantal gebruikers nog elk jaar groeit. Dit zal waarschijnlijk een behoorlijk deel van de stijging in advertentie-inkomsten voor zijn rekening nemen. Ter vergelijking: naar schatting van Piper Jaffray[1] zullen de inkomsten van de totale online advertentiemarkt in 2010 55 miljard dollar bedragen [M17]. Een aanzienlijk deel van de advertentie inkomsten van sociale netwerksites komt binnen bij Amerikaanse sites [M15].

Jaar	Advertentie-inkomsten
2007	$ 1.200.000.000
2008	$ 2.000.000.000
2009	$ 2.350.000.000
2010	$ 2.600.000.000
2011	$ 2.870.000.000
2012	$ 3.180.000.000
2013	$ 3.490.000.000

Tabel 1: (verwachte) advertentie inkomsten voor sociale netwerksites wereld wijd [M15] [M16]

Cijfers van de gezamelijke inkomsten van Nederlandse sociale netwerksites zijn niet voorhanden. De advertentie-inkomsten van de grootste sociale netwerksite van Nederland (Hyves) schat Emerce tussen de twee en vier miljoen euro [M18].

IDC[2] schat de advertentie-inkomsten voor sociale netwerksites in de VS gezamenlijk op vierhonderd miljoen dollar [M19]. In het onderzoek van Williamson [M15] wordt een bedrag genoemd van 350 miljoen dollar. Rich Greenfield[3] schatte in 2006 dat MySpace, de grootste Amerikaanse sociale netwerksite, dat jaar tweehonderd miljoen dollar aan advertentie-inkomsten zou ontvangen [M20].

Wanneer een sociale netwerksite miljoenen binnen krijgt voor het plaatsen van advertenties, wil dat nog niet zeggen dat men veel winst maakt. Hyves, dat gelanceerd werd in oktober 2004 en een jaar later begon advertenties te plaatsen, draaide pas sinds begin 2007 breakeven [M21].

[1] = Piper Jaffray is een Amerikaanse investeringsbank.
[2] = IDC is een Amerikaans bedrijf dat onderzoek doet naar IT, telecommunicatie en consumententechnologie.
[3] = Rich Greenfield is een analist van Pali Research

5.5.3 Aanpak plaatsing advertenties

Het maken van reclame op sociale netwerksites kan o.a. op de volgende manieren:

- Banners[1].
- Gebruikers een profielachtergrond met reclame laten maken in het kader van een wedstrijd.
- Grappige filmpjes met een commerciële boodschap [M22].
- Een profiel van het merk aanmaken[2] [M23]
- Door opinieleiders op de site voor het product te winnen [M22].
- Leden laten zien dat een van hun connecties het product gekocht heeft [M24] of waardeert [M25]
- Door een e-mail te versturen naar gebruikers van de site. [3]
- Door reclame te maken op het profiel van beroemdheden [M22].
- Opname in de nieuwsbrief van de sociale netwerksite.
- Door zich te mengen in discussies over het eigen product [M26] [M23].
- Leden van de site gratis proefmonsters aanbieden. [4]
- Het plaatsen van advertenties op profielen buiten de beheerders van de site om. [5]

1. Volgens RSMedia Sales, http://www.rsmediasales.nl/adverteren_op_girlsonly.php
2. Kan niet op alle sites maar is o.a. mogelijk op Hyves en Myspace
3. Kan niet op alle sites maar is o.a. mogelijk op Gurlz.nl, volgens de algemene regels van Gurlz.nl op 4 januari 2008 13:11, http://www.gurlz.nl/info.php?content=rules
4. Mexx heeft de vrouwelijke gebruikers van Hyves in 2007 een proefmonster van zijn nieuwe parfum aangeboden.
5. Op weblo.com kunnen gebruikers van bepaalde sociale netwerksites advertenties op hun profiel plaatsen en hier zelf geld voor ontvangen. Bij veel sites is dit in strijd met het gebruikersreglement.

Er zijn adviesfirma's die bedrijven helpen zichzelf te promoten op sociale netwerksites, voorbeelden hiervan zijn Deep Focus en New Delhi Digital Works.

5.5.4 Gericht adverteren

Het gericht plaatsen van advertenties is al op enkele sites mogelijk. Op Hyves heeft het gericht adverteren geleid tot een verdubbeling van het aantal muisklikken op advertenties [M27]. Leden van Facebook kregen[1] op hun homepage op de site advertenties te zien, afgestemd op hun interesses en bezigheden online. Wie bijvoorbeeld op Facebook meldde dat hij liefhebber is van een bepaalde band, kon reclame voor diens laatste cd te zien krijgen [M25]. Wanneer iemand een film een hoge waardering heeft gegeven, kan de adverteerder betalen om dit te communiceren aan de connecties van die persoon [M25]. Ook kunnen aankopen in webwinkels automatisch op het profiel geplaatst worden [M24]. Op Hyves kunnen adverteerders de doelgroep voor de advertentie selecteren op leeftijd, geslacht en woonplaats, maar niet op andere content op het profiel. Gericht adverteren gaat niet altijd goed. Een voorbeeld hiervan is het geval [M28] waarin op het Ajax-groepsprofiel een advertentie voor Feyenoord stond, wat leidde tot vijfhonderd boze emails.

Ik verwacht dat advertenties in de toekomst nog meer gericht worden op de inhoud van de profielen dan nu al gebeurd. Door middel het gebruik van associatieregels (dat in hoofdstuk 7, 8 en 9 aan de orde komt) kan ook geselecteerd worden op eigenschappen die een persoon waarschijnlijk heeft, maar niet op zijn profiel staan. Deze manier kan dus geschikt zijn om in de toekomst meer gericht te adverteren. Ook zie ik het gebruik van informatie die men ergens buiten het profiel heeft staan voor reclame op het profiel en vice versa binnen de mogelijkheden liggen. Een voorbeeld van deze integratie van informatie is (zoals al eerder genoemd) Facebook.

Nu worden advertenties op een profiel (wanneer er gericht geadverteerd wordt) toegespitst op het profiel zelf, hoewel een groot deel van de bezoekers van dat profiel andere eigenschappen zal hebben. Ik verwacht dat in de toekomst de advertenties die men ziet nadat men ingelogd heeft, meer toegespitst wordt op het profiel van de persoon die ingelogd heeft en hij aan hem gerichte advertenties ook ziet op profielen van anderen. De advertenties zouden dan aangepast worden aan de persoon die ze ziet, in plaats van de eigenschappen van die toebehoren aan de locatie van de advertentie, wat vaak het profiel van een ander is.

Volgens Post et al. [O20] boeken bedrijven het beste resultaat in het adverteren op sociale netwerksites door de gebruiker te faciliteren. Bedrijven dienen er voor zorgen dat zij een dialoog of dienst bieden aan de gebruiker en geen monoloog houden over hun eigen producten en diensten. Ook concluderen zij dat profiling een geschikte manier is om er voor te zorgen dat ongeveer 40 procent van de gebruikers op de banner klikt op de sociale netwerksite.

[1] Facebook kondigde in september 2009 aan te stoppen met dit omstreden advertentieprogramma.

5.5.2 Geen veilige omgeving

Volgens IDC [M29] zouden Amerikaanse sociale netwerken veel meer kunnen verdienen dan ze nu doen. Hoewel sites nu al adverteerders trekken, kan het lastig zijn adverteerders te overtuigen.

Eén van de obstakels van het adverteren op sociale netwerken is volgens IDC dat de sites geen omgeving kunnen bieden die veilig is voor een merk. Ook Schipperus[1] [M18] en een woordvoerder van de modeketen H&M [M23] geven aan dat het gebrek aan controle de reden is dat bedrijven niet enthousiast zijn over adverteren op sociale netwerksites.

Adverteerders willen hun advertenties namelijk niet weergegeven zien naast content die niet voldoet aan de sociale normen. Het dilemma voor sociale netwerksites wordt volgens hen nu of de sites de content gaan sturen, waardoor ze de populariteit zullen verliezen die de adverteerders naar de site trok [M29].

MySpace heeft speciaal voor adverteerders een apart gedeelte gemaakt content die het imago van adverteerders niet schaadt. In dat gedeelte kunnen geen tekst of foto's geplaatst worden van bijvoorbeeld: dronken jongeren, drugsgebruik of seksuele handelingen. Disney is een van de bedrijven die deze ruimte heeft gebruikt, voor de promotie van een film [M20].

Wal-Mart[2] is een voorbeeld van een bedrijf dat de gevolgen merkte van de mogelijkheid die gebruikers hebben hun eigen input te geven. Wal-Mart zette in augustus 2007 een promotiepagina op Facebook om aankopen voor studentenkamers te stimuleren. Binnen een paar weken stonden er meer negatieve berichten dan positieve berichten op het profiel. De negatieve berichten gingen vooral over het personeelsbeleid van Wal-Mart [M30].

5.5.3 Mening van gebruikers over advertenties

Ian Schafer[3] heeft gezegd dat het type gebruiker een nadeel is van adverteren op een sociale netwerksite. "Het is een cynisch publiek dat niet snel reageert op advertenties" [M31]. Dit blijkt ook uit een Brits onderzoek [O17] waarin 45 procent van de respondenten aangeeft het eens te zijn met de stelling: "Populaire community sites zijn geruïneerd door advertenties en zakelijke invloeden." In een onderzoek onder gebruikers van Hyves [O20], zegt vijfenvijftig procent van de respondenten geen interesse te hebben in aanbiedingen.

Nielsen [O19] geeft aan dat gewone banners bij gebruikers van sociale netwerksites minder goed werken. Omdat gebruikers zelf content aanleveren zouden zij volgens Nielsen [O19] een gevoel van eigenaarschap binnen de site hebben. Hij geeft aan dat daardoor gewone banners weinig effect hebben en adverteerders echt zullen moeten communiceren met gebruikers.

Volgens Nielsen [O19] komen sociale netwerksites in 2009 voor het dilemma te staan of zij gebruikers trekken met een strak design zonder veel advertenties, of kiezen voor de advertentie inkomsten waardoor de gebruikersgroei wel eens zou kunnen stagneren. Nielsen onderbouwd dit met de gebruikersgroei van het strak ogende Facebook en de afvlakkende gebruikersaantallen van MySpace, dat veel meer advertenties plaatst.

[1] = Michiel Schipperus werkt bij het Nederlandse internetbedrijf ISM eCompany
[2] = Wal-Mart is een Amerikaanse keten van warenhuizen
[3] = Ian Schafer is de CEO van Deep Focus, een interactief marketing bedrijf gevestigd in Amerika.

6 Informatie op sociale netwerksites waarheidsgetrouw?

Om de gevonden associatieregels beter te kunnen beoordelen is het goed om te weten in hoe verre de informatie waarvan zij zijn afgeleid waarheidsgetrouw is. Onbetrouwbaarheid van de data kan verschillende oorzaken hebben, namelijk een leugen, fout of datering. De verschillende achterliggende redenen om te liegen zijn in dit hoofdstuk uiteengezet, waarbij gekeken wordt of deze redenen ook aanwezig zijn bij foutieve informatie op sociale netwerksites.

6.1 Percentage leugens aan de hand van eigenschappen medium

Wanneer we weten onder welke omstandigheden veel gelogen wordt, kunnen we proberen aan de hand daarvan een inschatting te maken van het aantal leugens op profielen van sociale netwerksites.

In een onderzoek [O21] waarin studenten werd gevraagd hun sociale interactie en leugens te registreren, had men de hypothese dat het aantal leugens per medium in ieder geval afhankelijk is van de volgende eigenschappen van media:

- het synchroon verlopen van communicatie
    - Bij directe communicatie vinden meer leugens plaats, omdat zij meestal ongepland ontstaan.
- de automatische registratie van berichten
    - Wanneer berichten opgeslagen worden wordt er minder gelogen, waarschijnlijk uit angst om betrapt te worden.
- het gedistribueerd bevinden van de personen waartussen de communicatie plaatsvindt.
    - Als personen elkaar niet zien, wordt er meer gelogen, o.a. omdat niet alle leugens mogelijk zijn in fysieke aanwezigheid van de ander.

De data uit het onderzoek was consistent met de hypothese. De onderzoekers hebben zich niet gericht op een bepaald soort leugen.

Tabel 1 geeft de verschillende media afgezet tegen de drie hierboven genoemde eigenschappen uit het onderzoek aan. Echter in het onderzoek zijn sociale netwerksites niet meegenomen. Onderstaande tabel is gebaseerd op een tabel uit het onderzoek naar leugens in communicatie [O21]. Er zijn twee aanpassingen gedaan, namelijk het toevoegen van de kolom sociale netwerksites en het verwijderen van een dubbele negatie bij automatische registratie van berichten, die was ontstaan na vertaling van de tabel.

Eigenschappen:	**Face to Face**	**Telefoon**	**Instant Messaging**	**E-mail**	**Sociale netwerksites**
- Synchroon	X	X	X		
- Automatische registratie berichten			X	X	X
- Gedistribueerd		X	X	X	X
Percentage berichten dat minstens 1 leugen bevatte:	27 procent	37 procent	21 procent	14 procent	Onbekend

Tabel 2: Eigenschappen media en percentage berichten met leugen

Bij instant messaging (bijvoorbeeld door middel van MSN) worden berichten niet altijd automatisch opgeslagen, maar omdat de gesprekken geregistreerd worden voor de duur van het gesprek en eenvoudig opgeslagen kunnen worden voor langere tijd, hebben Hancock et al. besloten instant messaging te categoriseren als medium waarin de berichten automatisch geregistreerd worden.

Als we kijken naar de eigenschappen die door het eerder genoemde onderzoek [O21] als bepalend worden gezien voor het aantal leugens, komen sociale netwerksites overeen met email. Men ontvangt niet direct reactie op de berichten, de berichten worden automatisch opgeslagen en communicatie vindt gedistribueerd plaats. Daarom zou aangenomen kunnen worden dat het percentage leugens per bericht in de buurt ligt van het percentage leugens dat in e-mails gevonden kan worden.

Een verschil met e-mail is wel, dat de registratie van berichten wel automatisch gebeurt, maar door de gebruiker eenvoudig gewijzigd kan worden wanneer zij op het eigen profiel staan. Omdat leugens eenvoudig terug te draaien zijn, zal dit een gebruiker in mindere mate weerhouden een leugen te plaatsen.

E-mails zijn alleen meegenomen in het onderzoek [O21] wanneer het opstellen van de e-mail minstens tien minuten heeft gekost. Voor de andere drie vormen van media is een minimale gespreksduur van minstens tien minuten gesteld. Hierdoor kun je het percentage alleen vergelijken met profielen waarvan het opstellen minimaal tien minuten gekost heeft. Dit is waarschijnlijk lastig te bepalen aangezien gebruikers vaak slechts een paar gegevens per keer wijzigen of een kleine aanvulling doen.

Om het percentage leugens per profiel goed te kunnen bepalen, is aanvullend onderzoek nodig. Dit onderzoek alleen gehouden onder studenten, wat slechts met een beperkt deel van de sociale netwerksites vergelijking mogelijk maakt. Daarnaast hebben e-mails een meer tijdelijk karakter dan profielen.

6.3 Leugens

Wanneer een profiel oncorrecte informatie bevat kan dit meerdere oorzaken hebben. Eén daarvan is een bewuste leugen van de gebruiker. In deze paragraaf wordt onderzocht wat de (mogelijke) redenen zijn voor leugens op profielen. In de hierop volgende paragrafen komen overige oorzaken van fouten in profielen aan de orde.

In deze paragraaf worden redenen voor leugens in het algemeen genoemd. Aan de hand van literatuur over leugens op sociale netwerksites en andere online communicatie, wordt bekeken welke soorten leugens reden kunnen zijn voor fouten die aanwezig zijn in de data van sociale netwerksites. Het doel hiervan is een indruk te geven van eventuele oncorrecte informatie in de dataset van het experiment dat beschreven is in hoofdstuk 8 en 9.

6.3.1 Redenen voor leugens algemeen

Gamble identificeerde 4 types leugens [O22]:

- Sociale leugens: Leugens om iemand anders te helpen of beschermen.
    - Op sociale netwerksites geeft men vooral informatie over zichzelf, waardoor leugens om iemand te beschermen naar mijn inschatting weinig voor zullen komen.
- Zelfverbeterende leugens: Leugens om gezichtsverlies, schaamte, afwijzing en straf te voorkomen.
    - Deze leugens zullen naar mijn mening veel voorkomen, waarbij de voornaamste redenen voorkomen van gezichtsverlies en afwijzing zullen zijn.
- Zelfzuchtige leugens: Liegen om zichzelf te beschermen ten koste van anderen of om een misdaad te verbergen.
    - Zelfzuchtige leugens zullen vooral gebruikt worden door mensen die anderen lastig willen vallen of illegale activiteiten ontplooien op sociale netwerksites.
- Asociale leugens: Liegen om iemand anders opzettelijk pijn te doen of te kwetsen.
    - Deze leugens zullen net als zelfzuchtige leugens een minderheid vormen, maar ze komen zeker voor. Asociale leugens worden gebruikt bij identiteitsdiefstal of wanneer men iemand wil beledigen of kwetsen.

De Child Welfare League of America (CWLA) noemt de volgende redenen waarom kinderen liegen [M32]:

1. Om macht te krijgen
2. Om de grenzen te testen
3. Om autoriteit uit te dagen
4. Om iets te krijgen dat anders niet verkregen kan worden
5. Om wensen te vervullen
6. Om straf te vermijden
7. Om privacy te beschermen
8. Om zichzelf te beschermen tegen pijn of kwetsing
9. Om pijnlijke gevoelens of herinneringen te ontkennen
10. Om te bedreigingen, vernederingen of te voorkomen in een val te lopen
11. Om een ongemakkelijke situatie te voorkomen
12. Voor de lol
13. Om erbij te horen
14. Om te voorkomen dat vrienden in de problemen komen
15. Om de eigen status te verhogen
16. Om een onbedoelde fout te verbergen
17. Om belangrijker en leuker te lijken voor anderen
18. Om aan iemand zijn verwachtingen te voldoen
19. Om afwijzing te voorkomen
20. Om gebrek aan feitelijke informatie te compenseren

In de volgende paragraaf worden leugens die in artikelen genoemd zijn, ingedeeld in de categorieën van Gandal en gekoppeld aan redenen voor leugens opgesteld door de CWLA.

6.3.2 Redenen voor leugens op sociale netwerksites

De redenen die ik heb kunnen vinden in literatuur over sociale netwerksites zijn hieronder opgesomd. Wanneer deze reden zelf of de achterliggende reden voorkomt in de lijst redenen van de Cild welfare leage of America (CWLA), staat het nummer van de CWLA reden erachter.

Redenen voor leugens op sociale netwerksites	Type leugen volgens Gandal	Reden van CWLA	Paragraaf
Verkrijgen van toegang	Zelfverbeterend	4	6.3.3
Hokjesgeest / voldoen aan de norm	Zelfverbeterend	11, 13, 15, 17, 18, 19	6.3.4
Imago beschermen		10, 11, 19	6.3.5
Experimenteren met de identiteit	Zelfverbeterend[3]	4, 11, 12, 15, 17, 18, 19	6.3.6
Ongewilde aandacht en seksuele advances	Zelfverbeterend	8, 10, 11	6.3.7
Privacy beschermen	Zelfverbeterend	7	6.3.8
Beschermen tegen ouders en vreemden	Zelfverbeterend	6, 7, 8, 10,11	6.3.9
Illegale activiteiten verbergen	Zelfzuchtig	4, 6, 10, 14	6.3.10
Anderen lastig vallen	Asociale	2, 12	6.3.11
Imago beschadigen	Asociale	5, 12	6.3.12
Identificeren met een beroemdheid of groep	n.v.t.		6.3.13

Tabel 3: Leugens op sociale netwerksites gecategoriseerd

[3] Zelfonderzoekend zou beter passen, maar dit valt niet onder de types van Gandal. Zelfverbeterend is van de types van Gandal degene die het dichtst in de buurt komt.

6.3.3 Verkrijgen van toegang

Een deel van de sociale netwerksites bevat een leeftijdsrestrictie, meestal als gevolg van privacy-wetgeving. Kinderen onder een bepaalde leeftijd (de wettelijke leeftijdsgrens verschilt per land) kunnen hierdoor vaak hun eigen leeftijd niet invullen of krijgen geen toegang na het invullen van hun leeftijd. Wanneer alleen toegestane leeftijden ingevuld kunnen worden op het formulier, dwingen bedrijven jongere kinderen te liegen om toegang te krijgen [O2]. Om toch leeftijdsgenoten te kunnen vinden, draaien kinderen vaak de cijfers om, bijvoorbeeld 61 wanneer men 16 is. Een andere leeftijd die vaak voortkomt uit een leugen is 69, hier wordt voor gekozen omdat men dit humoristisch vindt [O5].

Over het algemeen is het eenvoudig te liegen over de leeftijd. De controle op de leeftijdsrestricties vanuit de overheid is in Nederland erg beperkt, het CPB treed pas op nadat er een klacht is ingediend [M33]. Veel Nederlandse sites gaan soepel om met de wettelijke leeftijdsrestricties. De minderheid legt de leeftijdsbeperkingen zo op dat kinderen een foutieve leeftijd moet invoeren om zich te kunnen aanmelden. Het grootste deel van de sites legt regels omtrend leeftijd (en evt toestemming van ouders) vast in het gebruikersreglement zonder dit te controleren (bijvoorbeeld Hyves) of legt geen enkele beperking of regel op (bijvoorbeeld Girlsonly). Bij sites die leeftijdsbeperkingen opleggen (of dit nu in voorwaarden is of in het fomulier) verschilt de leeftijdsgrens per site[1].

De controle vanuit de site beperkt zich meestal tot het controleren wat voor een leeftijd iemand in heeft gevoerd. Bij een beperkt aantal sites (waaronder Xanga.com [O23]) moet men bewijs leveren, bijvoorbeeld door een kopie van het paspoort op te sturen.

[1] Sugababes weigert jongeren onder de 13 jaar. Gurlz.nl en Hunkz.nl weigeren jongeren onder de 15 jaar. CU2 geeft aan dat jongeren onder de 18 jaar toestemming van hun ouders moeten vragen. (leeftijdgrenzen geldend in 2007)

6.3.4 Hokjesgeest / voldoen aan de norm

Sommige mensen voelen zich ongemakkelijk als ze zich in een hokje moeten plaatsen of hebben het gevoel zich anders voor te moeten doen om aan de heersende norm op de site te voldoen. Veel sites geven immers aan wat de norm is op de site door de opties in een specifieke volgorde te plaatsen [O2].

Anderen zijn van mening dat deze labels lang niet zo subtiel zijn dan die in het echte leven en zij geven aan dat dit stereotypering en leugens over persoonlijk informatie bevordert. Ook het niet willen voldoen aan de verplichting om bepaalde gegevens in te vullen kan leiden tot leugens [O2].

6.3.5 Imago beschermen

Personen kunnen schade ondervinden wanneer zij online gezien worden in een context die niet sociaal geaccepteerd wordt of niet past bij hun imago. Een voorbeeld is een Nederlandse politicus (Rouvoet, ChristenUnie), die negatief in het nieuws kwam toen bekend werd dat zijn dochter op haar profiel haar juf uitmaakte voor ‘kutwijf‘ en zich aan had gemeld bij een groepsprofiel van een andere politieke partij (de VVD) [M34].

Wanneer een medewerker beschuldigd wordt van sociaal onacceptabel gedrag op een sociale netwerksite [O4] of dingen op zijn profiel heeft staan waar zijn werkgever niet mee geassocieerd wil worden, kan het imago van het bedrijf hier schade van ondervinden wanneer bekend is dat de medewerker bij dat bedrijf werkt. Dit kan de reden zijn voor werkgevers richtlijnen op te stellen over het privé-gebruik van sociale netwerksites [M3].

Dergelijke richtlijnen of angst voor imagoschade kan leiden tot leugens. Hierbij kan men denken aan leugens over identificerende eigenschappen (om zich anoniemer te kunnen uiten) en leugens over interesses of eigenschappen die sterk van de norm of verwachting afwijken (waarmee eigenschappen die imagoschade kunnen veroorzaken verborgen worden).

6.3.6 Experimenteren met identiteit

Volgens Valkenburg et al. [O6] experimenteren jongeren op internet met hun identiteit om reacties van anderen op verschillende identiteitsuitingen te peilen, sociale contacten te verkrijgen en om over verlegenheid heen te komen. Zij kunnen op deze manier leren hoe zij zichzelf het beste kunnen presenteren en kunnen zij zonder weer te geven wie ze zijn eerst uitproberen hoe men reageert op bepaalde aspecten uit een persoon zijn identiteit. Iemand zou bijvoorbeeld eerst op een sociale netwerksite onder een fictieve identiteit kunnen weergeven dat hij homo is, dan kan hij zich aan de hand van de reacties voorbereiden op het moment dat hij aan vrienden gaat onthullen dat hij homo is. Maar ook kan men eigenschappen weergeven die men niet zelf bezit, bijvoorbeeld een ander geslacht aangeven, zo kan men volgens Bruckman [O24] ervaren hoe personen van de andere sekse behandeld en aangesproken worden. Ook presenteren mensen zich soms anders om zich beter voor te doen.

6.3.6 Ongewilde aandacht en seksuele avances

Volgens Bruckman [O24] zijn er vrouwen die zich op internet voordoen als man, vaak met de reden dat ze als vrouw ongewilde aandacht en seksuele avances krijgen, waardoor zij zich ongemakkelijk voelen. Ook noemt hij als reden voor vrouwen om zich voor te doen als man, dat vaak aangenomen wordt dat vrouwen hulp nodig hebben en de mannen seksuele gunsten verwachten in ruil voor technische hulp. Dit betekent niet dat zij zich een man voelen, maar dat ze zich zo voor doen met een bepaald doel [O2]. Het artikel van Bruckman [O24] gaat niet over sociale netwerksites, maar volgens Boyd [O5] vinden ongewenste avances zich ook plaats op sociale netwerksites. Zij zegt dat dit er ook toe kan leiden dat vrouwen liegen over hun relatiestatus. Het is mij niet bekend of vrouwen op sociale netwerksites ook als hulpbehoevend worden gezien en seksuele gunsten verwacht worden in ruil voor hulp.

6.3.8 Privacy beschermen

Persoonlijke informatie geplaatst op een profiel kan in een andere context privacygevoelige informatie worden, bijvoorbeeld wanneer verwacht wordt dat alleen vrienden het profiel bekijken, terwijl blijkt dat werkgevers en adverteerders, marketeers en journalisten ook op zoek gaan naar informatie op profielen [O2] [O5] [M35]. Ook verkopen veel sociale netwerksites gebruikersgegevens en -statistieken [O23].

Omdat men het bekijken van een profiel vanuit commercieel oogpunt ongewenst kan vinden, speelt privacy een rol [O4]. Door middel van het samenvoegen van informatie uit profielen van één persoon op verschillende sociale netwerksites, kan het beeld van een persoon steeds completer worden [O23]. In totaal kan men meer informatie ter besschikking krijgen dan de oorspronkelijke intentie van de gebruiker was.

Doormiddel van privacy wetgeving wordt men al enigszins beschermd. Kinderen genieten meer wettelijke privacy bescherming [O2] [M4], de leeftijdsgrenzen verschillen echter per land.
Zoals ook in paragraaf 5.1.2 aangegeven is, kun je er voor kiezen (een deel van de) gegevens afschermen voor onbekenden. Omdat de software van sociale netwerksites niet altijd waterdicht is [O13], brengt dit nog steeds een klein risico met zich mee.

6.3.9 Beschermen tegen ouders en vreemden

Bijna de helft van de ondervraagde HCC-leden[1], meent dat het grootste risico van gebruik van sociale netwerksites, is dat de veiligheid van jongeren in gevaar kan komen [M2]. Tieners voeren vaak valse identiteitsgegevens (als naam, leeftijd en woonplaats) in om zichzelf te beschermen. Ouders moedigen dit aan om de tieners te beschermen tegen vreemden. Veel tieners vullen de valse informatie echter in om zichzelf te beschermen tegen hun nieuwsgierige en controlerende ouders [O3]. Volgens het Sociaal Cultureel Planbureau hebben veel tieners persoonlijke informatie en foto's op een sociale netwerksite staan zonder dat hun ouders daarvan weten [O25].

6.3.10 Illegale activiteiten verbergen

Sociale netwerksites worden gebruikt voor illegale activiteiten, waaronder drugshandel [O5]. Mensen die communiceren over illegale zaken, als drugshandel of -gebruik, zullen dit meestal niet onder hun eigen naam doen [O4].

Ook kan het hebben van bepaalde eigenschappen een voordeel of voorwaarde zijn, voor het uitvoeren van bepaalde illegale activiteiten. Pedofielen komen o.a. via sociale netwerksites in contact met kinderen [M6]. Om makkelijker met hen in contact te komen, doen deze volwassenen zich vaak anders voor. De informatie die zij in hun profiel hebben staan klopt vaak (gedeeltelijk) niet. Een voorbeeld hiervan is een man die zich voordeed als vrouwelijk model en modellenscout [M7] en minderjarige meisjes onder druk zette ontuchtige handelingen uit te voeren.

6.3.11 Anderen lastig vallen

Wanneer men anderen wil lastig vallen kan men een andere (al dan niet fictieve) identiteit aannemen om te voorkomen dat men last heeft van de consequenties van het gedrag of om toegang te verkrijgen tot groepspagina's of groepsfora. Voorbeelden van het lastig vallen op sociale netwerksites zijn o.a. trolling[1] en online pesten.

Om anderen lastig te vallen of voor de gek te houden, kan ook een specifieke identiteit aangenomen worden, bijvoorbeeld die van een beroemdheid. Dat identiteitsfraude van beroemdheden voorkomt is bekend [M36] [M37] en wordt in paragraaf 6.3.12 beschreven. Of het lastig vallen van anderen bij deze gevallen de reden was voor het aannemen van die identiteit is echter niet duidelijk.

[1] Trolling is het vertrouwen winnen van anderen, waarna men probeert een discussie een andere wending te geven, ruzie of commotie te veroorzaken of bewust verkeerd advies te geven.

6.3.12 Imago beschadigen

Wanneer men een andere (niet fictieve) identiteit aanneemt, is er sprake van identiteitsdiefstal. Als iemand zich voordoet als een ander kan deze persoon het imago of de reputatie van die persoon schade toedoen. Het aannemen van de identiteit van een ander is online (in vergelijking met de fysieke wereld) relatief eenvoudig, aangezien er een beperkt aantal eigenschappen van de identiteit wordt weergegeven [O4]. Dat er daadwerkelijk gevallen zijn van identiteitsdiefstal op sociale netwerksites blijkt uit nieuwsberichten [M37] [M36].

Het is niet duidelijk of het beschadigen van iemands imago over het algemeen het doel is van identiteitsdiefstal op sociale netwerksites. Het ervaren hoe het is om beroemd te zijn (experimenteren met de identiteit) of lastig vallen van anderen (fans) kan hier ook reden voor zijn. Echter, ook wanneer het aanbrengen van imagoschade niet het doel was, zal dit vaak wel een gevolg zijn van identiteitsdiefstal.

6.3.13 Identificeren met een beroemdheid of groep

In sommige gevallen is de kans groter dat het profiel wordt beheerd door een ander. Dit gebeurd vaak bij profielen van:

- Beroemdheden
- Tekenfilmkarakters
- Objecten
- Groepen
- Organisaties

Bij beroemdheden of tekenfilmkarakters is deze persoon vaak een fan [O26]. Om aan te geven dat men fan is maakt men een connectie aan met dit profiel. Op grotere sociale netwerksites wordt het profiel vaker beheerd door de beroemdheid of zijn werknemers. Een beroemdheid gebruikt het bijvoorbeeld om zijn fans op de hoogte te houden. Karakters kunnen beheerd worden door het bedrijf dat het karakter exploiteert.

Bij objecten is de het onduidelijker waarom iemand hier een profiel van aanmaakt, bijvoorbeeld profielen van zout en peper [O26].

Groepen en organisaties kunnen bijvoorbeeld de universiteit, sportclub of werkgever zijn. Bij een deel van de netwerksites worden profielen die niet over een persoon gaan verwijderd om wildgroei tegen te gaan [O27]. Op andere sociale netwerksites is dit toegestaan en gereguleerd door een speciaal profiel voor organisaties of groepen te maken. Hyves heeft bijvoorbeeld een profielsoort voor groepen, families en verenigingen.

6.4 Redenen om niet te liegen

Ik verwacht dat een groot deel van de gebruikers die niet liegt dit doet omdat zij hier geen noodzaak toe zien of het gevoel hebben niks te verbergen te hebben. Echter er zijn ook mensen die er bewust voor kiezen niet te liegen. Redenen kunnen dan o.a. zijn [O2] [M38]:

- Realistischer maken van interactie
- Persoonlijke informatie geeft aanleiding tot gesprek
- Betere gebruikerservaring door gerichte advertenties
- Eenvoudiger gevonden kunnen worden door (oude) bekenden
- Informatie op profiel kan invloed hebben op sollicitatie
- Uit angst voor sancties van de beheerder van de site of justitie
- Vrienden die de waarheid kennen gaan vragen stellen

Hieronder worden deze redenen beschreven.

6.4.1 Realistischer maken van interactie

Wanneer een gebruiker op een sociale netwerksite een formulier in moet vullen voor zijn profiel, stereotypeert de gebruiker zichzelf bewust of onbewust. Voorstanders van deze classificatie zien hier het voordeel in dat informatie die aanwezig is bij fysieke interactie nu ook aanwezig is op de site [O2]. Het realistischer maken van de interactie, kan dus een reden zijn een waarheidsgetrouw profiel te plaatsen.

6.4.2 Persoonlijke informatie geeft aanleiding tot gesprek

Het delen van persoonlijke informatie (bijvoorbeeld de woonplaats) kan aanleiding zijn voor gesprekken met onbekenden. Volgens Boyd [O2] leidt informatie over leeftijd, geslacht, woonplaats e.d. zelden tot diepe gesprekken, maar vormen deze onderwerpen wel een makkelijke ingang voor een gesprek.

6.4.3 Betere gebruikerservaring door gerichte advertenties

De informatie op profielen wordt gebruikt door derden, bijvoorbeeld adverteerders. Bedrijven geven als argument voor het verzamelen van persoonlijke informatie, dat dit zorgt voor een betere gebruikerservaring omdat gebruikers gerichte advertenties ontvangen [O2]. Of er gebruikers zijn die omwille van deze reden er voor kiezen niet te liegen is mij echter niet bekend.

6.4.4 Eenvoudiger gevonden kunnen worden door (oude) bekenden

Daarnaast kan het zijn dat bekenden een persoon niet kunnen vinden op een sociaal netwerk omdat hij foutieve informatie ingevoerd heeft. Dit zal in sommige gevallen de bedoeling zijn (redenen hiervoor zijn eerder in dit hoofdstuk uiteengezet), maar kan ook een nadeel zijn omdat een sociaal netwerk nu eenmaal draait om sociale contacten [O2].

6.4.5 Informatie op profiel kan invloed hebben op sollicitatie

Werkgevers kunnen via professionele netwerksites als LinkedIn, de juiste persoon vinden voor een functie [O15]. Wanneer een sollicitant gevonden is (al dan niet via een sociale netwerksite), gaat een deel van de werkgevers op zoek naar meer informatie op internet. Eén op de vijf werkgevers vind informatie over de sollicitant op het internet waarmee de werkgever nog niet bekend was [M38]. Ik neem aan dat een deel hiervan afkomstig is van sociale netwerksites. Voor 59 procent van de werkgevers die informatie vonden, heeft de online reputatie invloed gehad op de beslissing iemand wel of niet aan te nemen [M38]. Meer dan de helft van de volwassenen vinden dit onethisch [O18], maar sollicitanten zouden er ook gebruik van kunnen maken door die informatie op hun profiel te zetten waarmee zij een betere en meer complete indruk geven dan andere sollicitanten. Dertien procent van de HR managers geeft aan dat zij informatie online hebben gevonden waardoor de sollicitant aangenomen werd, terwijl zij de persoon zonder die informatie niet aangenomen hadden [M38].

6.4.6 Vrienden die de waarheid kennen gaan vragen stellen

Ook kunnen vrienden vragen gaan stellen over de ingevulde gegevens wanneer zij weten dat dit niet klopt [O5]. Vrienden oefenen een soort sociale controle uit. Daarom zal een profiel op een sociale netwerksite waarschijnlijk minder fouten bevatten dan een profiel op bijvoorbeeld een datingsite (waar deze sociale controle minder heerst).

6.4.7 Uit angst voor sancties van de beheerder van de site of justitie

Wanneer een gebruiker zijn profiel niet naar realiteit invult is de kans op sancties van de beheerder erg klein. Allereerst moet de beheerder erachter komen. Vaak worden profielen niet gemonitord, maar zullen beheerders een overtreding opmerken door een klacht van een andere gebruiker (of diens ouders). Wanneer de beheerder de leugen heeft opgemerkt kan deze optreden door de gebruiker erop aan te spreken, het account te verwijderen, het ip-adres toegang te ontzeggen of aangifte te doen (dit laatste alleen bij wetsovertredingen, bijvoorbeeld identiteitsfraude of kinderlokkerij). Veel sociale netwerksites (waaronder Facebook en Hyves) geven in hun gebruikersregelement het profiel te verwijderen wanneer gebruikers zich niet aan de regels houden. Er worden geen cijfers naar buiten gebracht over het aantal profielen dat verwijderd wordt naar aanleiding van het overschrijden van de regels.

6.5 Wel foutief, geen leugen

Profielen zijn vanuit een bepaald standpunt geschreven en kunnen hierdoor verkeerd geïnterpreteerd worden. In sommige gevallen bevatten profielen ook ongewild misleidende/foutieve informatie [O5]. Verouderde of per ongeluk foutief ingevoerde informatie leidt tot onbetrouwbaardere data op de profielen.

Per ongeluk foutief ingevoerd
Het maken van een typfout is menselijk, hierdoor komen er onbedoeld fouten in het profiel te staan. Omdat er sociale controle op sociale netwerksites heerst [O5], kunnen contacten de eigenaar van het profiel o.a. op onbewuste fouten wijzen. Dit maakt de eigenaar bewust van de foute informatie.

Verouderde data
Wanneer de gebruiker zijn profiel niet meer update, bevat dit naar verloop van tijd incorrecte informatie. Op het merendeel van de sociale netwerksites is een profiel niet of nauwelijks te verwijderen [M39].
Passieve gebruikers loggen hoogstens af en toe nog in om vriendaanvragen te bevestigen of berichten te beantwoorden, maar wijzigen hun profieldata niet. Zonder nieuwe informatie, worden de profielen een statische weergave van een verouderde representatie van de persoon [O5].

6.6 Meerdere identiteiten

In de fysieke wereld is het gebruikelijk dat mensen zich anders presenteren (een andere identiteit aannemen) bij een andere gelegenheid, zonder dat een van deze identiteiten niet klopt. Het is niet ongebruikelijk voor mensen om meerdere e-mailadressen of telefoonnummers te hebben als manier om controle te houden over de toegang tot de verschillende identiteiten. Als gevolg hiervan maken zij ook online meerdere accounts aan [O2].

Dat er mensen zijn die online meerdere identiteiten aanmaken blijkt ook uit het gebruik van o.a. Poken, Onxiam.com, OpenId.net en OpenSocial.org. Op Onxiam kan iemand links naar zijn identiteiten toevoegen, beheren en laten zien aan anderen. De poken is een gadget waarmee men het adres van alle profielen die men online heeft in een keer met een ander kan uitwisselen [M41]. Door middel van Openid en OpenSocial worden gegevens tussen verschillende sociale netwerksites uitgewisseld.

Wanneer iemand meerdere identiteiten heeft op één sociale netwerksite, vertroebeld dit de verzameling profielen op de site omdat de ene persoon dan meer identiteiten zal hebben als de ander. Statistieken op de site zullen hierdoor negatief beïnvloed worden.

Om leugens te voorkomen, willen bedrijven dat gebruikers universele profielen aanmaken die voor meerdere websites gelden. Een voorbeeld hiervan is Microsoft, met zijn .Net passport initiatief. Zij suggereren hiermee dat ze het bewegen op internet eenvoudiger maken voor de gebruiker omdat hij zich makkelijker aan kan melden zonder elke keer opnieuw zijn gegevens in te hoeven voeren. De meeste gebruikers hebben echter geen interesse in het samenvoegen van al hun identiteiten tot één identiteit. Boyd [O2] geeft aan dat universele accounts leugens niet zullen voorkomen, maar gebruikers dwingen meerdere accounts aan te maken of overal te liegen.

6.7 Gevolgen van foutieve informatie

In deze paragraaf zijn gevolgen van foutieve informatie voor adverteerders en voor dit onderzoek uiteengezet.

6.7.2 Gevolgen voor adverteerders

Foutieve informatie op iemand zijn profiel heeft negatieve gevolgen voor de adverteerders.

Wanneer informatie op profielen niet correct is, heeft de beheerder een minder goed beeld van zijn gebruikers. Hierdoor zal het moeilijker zijn gerichte advertenties te plaatsen. De adverteerder is natuurlijk niet bij gebaad bij het niet kunnen plaatsen van gerichte advertenties omdat hij niet de juiste doelgroep bereikt.

De beheerder van een site kan rekening houden met verouderde informatie door de datum waarop er ingelogd of geüpdate is bij te houden. Dit geeft echter slechts een indicatie; men weet niet of de gegevens in de tussentijd gewijzigd zijn. Aan de hand van de update-datum zou men zich bij het gebruik van de data (bijvoorbeeld voor statistieken) kunnen focussen op gebruikers die informatie recentelijk gewijzigd hebben. Vanzelf sprekend is het ook afhankelijk van het type gegevens of deze snel verouderd. In dit onderzoek is de wijziging of invoerdatum van de gegevens niet meegenomen omdat deze data niet voorhanden was.

6.7.3 Gevolgen foutieve informatie voor dit onderzoek

Een onderzoek dat gebruik maakt van onbetrouwbare input, geeft geen betrouwbare onderzoeksresultaten. De data van profielen vormen input voor dit onderzoek. Wanneer een deel van deze informatie niet klopt, is de output minder betrouwbaar. De output is immers gebaseerd op de realiteit die door de gebruikers wordt gepresenteerd, welke niet per definitie correct is.

Het gebruik van meerdere identiteiten door één persoon maakt de data minder betrouwbaar. Men maakt namelijk een keuze wat men in welk profiel kenbaar maakt. Hierdoor ontstaat een onvolledig beeld van iemands persoonlijkheid. Een dergelijk onvolledig beeld ontstaat ook wanneer iemand niet alle velden invult. Echter wanneer een persoon meerdere identiteiten heeft (aangenomen dat deze identiteiten geen foutieve data bevatten), heeft dit wel een andere invloed op de resultaten van dit onderzoek dan iemand die niet alle velden invoert.

Wanneer mensen meerdere identiteiten (profielen) hebben op één sociale netwerksite, zal de ene persoon waarschijnlijk meer identiteiten hebben dan de ander. Iemand met veel identiteiten heeft een grotere invloed op de resultaten dan iemand met weinig identiteiten, aangezien hij meerdere keren in de onderzoeksdata voorkomt. Dit maakt de kwaliteitsberekeningen van de regels onbetrouwbaarder.

De aanpak die gekozen is in dit onderzoek met betrekking tot onbetrouwbare data is beschreven in hoofdstuk 7.

7 Aanpak experiment: betrouwbaarheid van data

In dit hoofdstuk wordt beschreven hoe het experiment naar betrouwbaarheid van data op sociale netwerksites is uitgevoerd.

7.1 Onderzochte sociale netwerksites

Er is contact opgenomen met veertig sociale netwerksites met de vraag of zij medewerking wilden verlenen. Drie sites hebben hun medewerking toegezegd aan het databetrouwbaarheidsonderzoek: Ikku.nl, Hunkz.nl/Gurlz.nl en Girlsonly.nl. Hunkz.nl en Gurlz.nl zijn twee verschillende internetadressen, maar de twee sites vormen één community. Het enige verschil is het kleurgebruik op de sites en de site waar men zich afhankelijk van het geslacht bij aanmeld.

Uit hoofdstuk 4 blijkt dat niet alle data op een sociale netwerksite betrouwbaar is en er meerdere oorzaken kunnen zijn voor onbetrouwbare data. De gegevens die onbetrouwbaar zijn, zijn in sommige gevallen gerelateerd aan de oorzaak. Bijvoorbeeld wanneer een kind liegt over zijn leeftijd omdat hij anders geen toegang krijgt tot de site, of wanneer een vrouw zich voordoet als een man om seksuele avances te voorkomen. Wanneer de data niet geüpdate is, zal dit meestal geen verband hebben met de veranderde data. Sommige soorten gegevens veranderen echter vaker als andere, geboortedatum is een constant attribuut, maar een woonplaats kan wijzigen.

Dat niet alle data betrouwbaar is en dat er over verschillende soorten gegevens gelogen kan zijn, staat beschreven in hoofdstuk 4. Om de betrouwbaarheid van de resultaten van het datamining experiment in te kunnen schatten, is het van belang informatie te hebben over de databetrouwbaarheid van de informatie op de profielen. De databetrouwbaarheid van de informatie is onderzocht middels een enquête gehouden onder gebruikers van de sociale netwerksites die hun medewerking hebben verleend aan dit onderzoek.

Lui en Maes [O27] geven aan dat het analyseren van gegevens van meerdere sociale netwerksites als voordeel heeft dat de gebruikte site minder invloed heeft op de uitkomst van het onderzoek. Als nadeel geven zij aan dat er een overlap is wat betreft gebruikers. Dit heeft als gevolg dat gebruikers die lid zijn van meerdere onderzochte sites, meer invloed hebben op de uiteindelijke resultaten.

In het onderzoek van Lui en Maes [O27] wordt de overlap geschat op vijftien procent. Dit percentage is naar mijn mening aardig hoog voor sites met een klein marktaandeel. In hun artikel geven zij niet aan welke sociale netwerksites zij gebruikt hebben in hun onderzoek. In paragraaf 5.4.3 staat dat het overgrote deel van de gebruikers van een specifieke sociale netwerksite uit één of enkele landen afkomstig zijn. Wanneer Lui en Maes grote sociale netwerksites gebruiken waarbij de geografische verdeling van de leden redelijk overeenkomt, zal de overlap tussen twee sociale netwerksites relatief groot zijn. Bij een kleine geografische overlap of een lager gebruikersaantal per site, zal de overlap in gebruikers kleiner zijn.

De sites die medewerking verlenen aan dit onderzoek opereren wel in het zelfde geografische gebied, ze hebben allen vooral Nederlandse leden. Het zijn in vergelijking met de grootste Nederlandse sociale netwerksite (Hyves, ruim 7 miljoen Nederlandse leden in december 2007 [M40]) erg kleine sites. De sites die medewerking hebben verleend aan dit onderzoek hebben beperkte gebruikersaantallen, waardoor maar een klein deel van de totale gebruikersgroep van sociale netwerksites lid is van een van deze sites. De kans is dus relatief klein dat gebruikers bij twee van de onderzochte sites staan ingeschreven. Ik acht de kans groter dat mensen meerdere profielen bij één van de medewerkende sites hebben. Toch dient er wel rekening mee te worden gehouden dat er mensen zijn die meer invloed hebben op het resultaten van het onderzoek, doordat zij lid zijn van meerdere medewerkende sites of meerdere profielen bij één site hebben.

7.2 Aanpak Enquête

In deze paragraaf wordt beschreven wie de onderzoeksobjecten zijn en hoe zij benaderd zijn. Het gewenste aantal respondenten en de acties ondernomen om voldoende respons te verkrijgen worden genoemd. Ook is uiteengezet waarom sociale wenselijkheidsvragen zijn gebruikt om de betrouwbaarheid van het experiment te vergroten.

7.2.1 Onderzoeksobjecten

Het doel was minstens honderd respondenten per site te hebben (Hunkz en Gurlz als één site gerekend). De leden van de sites die hun medewerking toegezegd hebben zijn benaderd. Van Hunkz, Gurlz en Girlsonly zijn alle leden benaderd. Van Ikku.nl zijn in eerste instantie de drieduizend leden die het laatst ingelogd hebben benaderd. De leden van deze vier sites zijn benaderd door middel van een e-mail. De leden waren op deze manier makkelijk te bereiken omdat de emailadressen bij de beheerder van de site bekend zijn (in tegenstelling tot bijvoorbeeld het adres of telefoonnummer). De e-mail is gestuurd door de beheerders van de sites. Er is voor gekozen de leden via de beheerders van de sites te benaderen om er voor te zorgen dat zij de e-mail vertrouwen en om de privacy van de leden te beschermen.

De leden konden de enquête (welke te vinden is in de bijlagen van dit document) anoniem invullen. De mogelijkheid bestaat dat leden de enquête meerdere malen invullen, aangezien deze anoniem is. Echter wanneer zij dit doen om de kans te vergroten de prijs te winnen, zal dit opgemerkt worden wanneer zij dit doen onder hetzelfde emailadres. Aangezien mensen maar een beperkt aantal emailadressen hebben, zal het effect op de resultaten beperkt blijven. Hierbij neem ik aan dat men voor een beperkte kans op een prijs van slechts vijftig euro niet een nieuw emailadres aanmaakt.

De duur van de enquête was voor de aanvang van de enquête vastgesteld op vijftien dagen. De reden hiervoor is dat de onderzoeksobjecten dan voldoende tijd hebben gehad de enquête te beantwoorden. Omdat de oproep van de enquête via email verstuurd is schat ik de kans dat zij dit na deze periode nog doen klein in.

Na een duur van twaalf dagen had ik bij een deel van de sites nog niet voldoende respons. Om de respons te verhogen van de sites met een te lage respons, zijn verschillende acties ondernomen. Er is een herinneringsmail gestuurd waarin vermeld werd dat de deadline voor het invullen van de enquête en het winnen van de prijs een week verschoven werd. Ook is het aantal benaderde (Ikku-)leden uitgebreid en is er op een van de sites een bericht met een link naar de enquête op de Hunkz en Gurlz home pagina geplaatst.

7.2.2 Wedstrijd

Om de 100 respondenten per site te kunnen bereiken, is een prijs uitgeloofd, namelijk een tijdschriftencadeaubon ter waarde van vijftig euro. De enquête was anoniem maar als men de prijs wilde winnen diende men zijn e-mailadres in te voeren. Men kon er echter ook voor kiezen het e-mailadres niet in te vullen om de vragenlijst anoniem te houden.

Volgens Aadahl en Jørgensen [O28] zou het verloten van een prijs geen invloed hebben op de antwoorden en ook nauwelijks op de respons, mensen zullen alleen sneller antwoorden. Het onderzoek van Aadahl en Jørgensen [O28] is echter gehouden onder een oudere leeftijdsgroep en gaat het om een enquête via de post. Ik denk dat jongeren eerder geneigd zijn te antwoorden wanneer er een prijs wordt verloot, omdat zij over het algemeen meer tijd te besteden hebben en over minder geld beschikken om zelf de prijs aan te schaffen die verloot wordt. Ook denk ik dat een enquête via de post langer bewaard wordt met de intentie deze in te vullen. Als een e-mail eenmaal ouder is dan een week wordt deze vaak niet meer bekeken is mijn ervaring, ook al had de persoon wel die intentie te reageren. Daarom vermoedt ik dat een snellere respons in dit geval wel degelijk invloed heeft op het aantal respondenten.

7.2.3 Sociale wenselijkheid vragenlijst

Sociale wenselijkheid kan volgens Ganster et al. tot foutieve resultaten leiden of echte resultaten onderdrukken [O29]. Er is door de jaren heen veel onderzoek gedaan naar sociale wenselijkheid in relatie tot andere eigenschappen. Onderwerpen waarop een hoog aantal sociaal wenselijke antwoorden wordt gegeven zijn bijvoorbeeld roken, alcoholgebruik, seksueel gedrag, lengte, gewicht, gebruik van de veiligheidsgordel, politieke voorkeur en afstanden tot de dichtstbijzijnde supermarkt. [O30]

Omdat sociaal wenselijke respons een enquête minder betrouwbaar maakt, is besloten binnen dit onderzoek hierop in te spelen. Dit is gedaan door te kiezen voor een enquête via internet, de mogelijkheid de enquête anoniem in te vullen en door sociale wenselijkheidsvragen in de enquête te verwerken. Uit onderzoek van Joinson [O31] is gebleken dat enquêtes die anoniem via internet ingevuld worden significant lagere sociale wenselijkheid respons vertoont dan niet anonieme enquêtes via de post. De keuze voor email had ook een praktische reden, zoals beschreven is in paragraaf 7.2.1.

In paragraaf 7.3.7 staat beschreven hoe het risico op het geven van sociaal wenselijke antwoorden verkleind is door middel van een aantal vragen.

7.2.4 Interpretatie vragenlijst

Om te testen of de vragenlijst juist geïnterpreteerd wordt, heb ik tachtig mensen gevraagd deze vragenlijst in te vullen en feedback te geven op de enquête. Vijfendertig mensen hebben de vragenlijst geheel of gedeeltelijk ingevuld. Deze groep bestond uit 29 mannen, 3 vrouwen en 3 mensen die hun geslacht niet hebben ingevuld. De leeftijd van de respondenten was gemiddeld 23,8 met 23 als mediaan. De groep bestond uit 25 HBO-ers, 5 WO-ers en 5 personen die hun opleiding niet hebben ingevuld. De verwachting was dat deze groep niet geheel representatief was ten opzichte van de leden van de meewerkende sociale netwerksites. Voor deze groep is gekozen omdat zij eenvoudig bereikbaar waren en de enquête met deze groep achteraf nabesproken kon worden. Omdat deze groep niet geheel representatief is voor de onderzoeksobjecten, maakt deze test de vragenlijst niet waterdicht.

Met behulp van de test kwamen er echter wel een aantal onduidelijkheden uit de vragenlijst naar boven. Twee opmerkingen kwamen herhaaldelijk naar voren. Ten eerste gaven de sociale wenselijkheidsvragen een vreemde indruk. Deze indruk werd veroorzaakt omdat de vragen in onderwerp afwijken van de rest van de vragenlijst en niet aangegeven is waarom deze vragen gesteld werden. Ten tweede werd opgemerkt dat de vraag over leeftijd beter anders gesteld zou kunnen worden, dit was in eerste instantie een meerkeuzevraag. De opmerking over de sociale wenselijkheidsvragen was verwacht, aangezien het doel van deze vragen niet genoemd is omdat dit de antwoorden zou kunnen beïnvloeden. Deze klacht is dus niet weggenomen. De overige opmerkingen zijn gebruikt om de vragenlijst te verbeteren.

7.2.5 Betrouwbaarheid experiment

Omdat de enquête een aantal vragen bevat die gevoelig kunnen liggen, kan dit een reden zijn voor respondenten om te liegen bij de beantwoording van de vragenlijst. Vragen die gevoelig kunnen liggen zijn die waarin gevraagd wordt of gegevens kloppen. Wanneer gegevens op het profiel niet kloppen, zal dit vaak komen doordat men gelogen heeft. Omdat dit een sociaal onwenselijk antwoord is, kan het voorkomen dat iemand aangeeft dat de gegevens kloppen, omdat hij zijn leugen niet kenbaar wil maken. Weliswaar zijn er ook andere redenen voor foutieve gegevens, bijvoorbeeld dat deze niet geüpdate zijn. Ik verwacht dat een respondent eerder aan een leugen zal denken dan aan een fout op zijn profiel, omdat de leugen bewust gemaakt is. Wanneer een respondent dagelijkse leven geneigd is sociaal wenselijke antwoorden te geven, zal hij waarschijnlijk in deze enquête eerder geneigd zijn te liegen.

7.3 Inhoud vragenlijst

De vragenlijsten die gebruikt zijn, staan in de bijlagen. Er is per site een vragenlijst gemaakt. De vragenlijsten bevatten enkele kleine verschillen, meestal als gevolg van verschillen tussen de verschillende sites.

Er zijn vragen gesteld over de volgende onderwerpen:

- Meerdere identiteiten (alleen op de betreffende site)
- Betrouwbaarheid standaardgegevens op het profiel
- Vragenlijst op het profiel
- Betrouwbaarheid overige gegevens
- Reden foutieve gegevens
- Persoonsgegevens (leeftijd, geslacht, opleidingsniveau)
- Sociale wenselijkheid
- Betrouwbaarheid profielen van vrienden
- Deelname aan wedstrijd

7.3.1 Meerdere identiteiten

Als mensen meerdere identiteiten hebben waarvan een op de onderzoekssite en daarnaast één of meerdere op andere sites dan is dit voor dit onderzoek niet van belang. Daarom is alleen gevraagd hoeveel profielen men op de betreffende sociale netwerksite heeft. Ook werd het verschil tussen de verschillende identiteiten gevraagd en de reden van het bezit van meerdere identiteiten. De vraag naar het aantal identiteiten was verplicht, maar de vraag om de reden voor meerdere identiteiten en de verschillen tussen de identiteiten waren niet verplicht. Deze twee vragen konden niet verplicht worden gesteld omdat deze niet beantwoord hoeven worden als men één profiel heeft. De gebruikte site die gebruikt is voor het voorleggen van de enquête, beschikte niet over de mogelijkheid door te vragen afhankelijk van het antwoord op de voorgaande vraag. Daarnaast is de reden voor het hebben van meerdere identiteiten voor dit onderzoek ook niet belangrijk, het levert alleen interessante achtergrondinformatie.

Bij het vragen naar het aantal profielen per site werd Hunkz en Gurlz als één site gezien omdat het één community is. Bij Hunkz en Gurlz werd ook gevraagd op welke van de twee sites men het profiel of de profielen heeft staan, of dat men bij beide sites één of meerdere profielen heeft.

7.3.2 Betrouwbaarheid standaardgegevens op het profiel

De betrouwbaarheid van standaardgegevens op het profiel werd onderzocht door van twaalf (op deze sites) veelgenoemde eigenschappen te vragen of de eigenschap klopt met de werkelijkheid. Deze worden standaardgegevens genoemd hoewel zij niet per definitie in elk profiel voorkomen.

Per site verschilde of er wel of geen veld was voor elk van de standaardgegevens, ook was dit veld in sommige gevallen wel en in andere gevallen niet verplicht. Bij ontbreken van een veld voor de eigenschap, kan de gebruiker deze eigenschap in de vrije ruimte genoemd hebben. Omdat de gebruiker bij een niet verplicht veld of het ontbreken van het veld voor de eigenschap, de mogelijkheid heeft de eigenschap niet op zijn profiel te zetten werd bij de vragen van de enquête over de betrouwbaarheid van de gegevens ook de optie 'Niet ingevuld' gegeven.

Omdat de vragen naar betrouwbaarheid van groot belang zijn voor dit databetrouwbaarheidsonderzoek en iedereen deze vragen kan invullen (mede door de optie 'niet ingevuld'), zijn deze vragen verplicht.

Voor de volgende gegevens is gevraagd of zij klopten:

- Leeftijd
- Geslacht
- Provincie
- Woonplaats
- Opleiding
- Relatiestatus
- Geaardheid
- Huidskleur
- Kleur ogen
- Andere uiterlijke kenmerken
- Hobby's
- Sterrenbeeld

Wanneer mensen meer dan één profiel op de betreffende sociale netwerksite hadden, werd hun gevraagd de betrouwbaarheidsvragen te beantwoorden voor het laatst bewerkte profiel. Er is gevraagd het laatst bewerkte profiel te gebruiken om te voorkomen dat de onderzoeksobjecten zelf de keuze maken en omdat de onderzoeksobjecten vragen hierover waarschijnlijk beter kunnen beantwoorden.

Wanneer de onderzoeksobjecten vrij gelaten worden in de keuze van het profiel, kunnen zij deze mogelijk baseren op de vragen en hierdoor kiezen voor het profiel dat het meest overeenkomt met de werkelijkheid. Een keuze voor het sociaal wenselijke profiel, maakt de totale selectie van besproken profielen van alle onderzoeksobjecten tot een minder goede steekproef.

De onderzoeksobjecten kunnen vragen over hun meest recente profiel waarschijnlijk beter beantwoorden omdat het aannemelijk is dat zij eigenschappen van dit profiel meer uit hun hoofd zullen kennen. Ik verwachtte dat het grootste deel van de respondenten voor het beantwoorden van deze vragenlijst niet hun profiel te voorschijn zouden halen om dit te controleren.

De keuze voor het laatst bewerkte profiel kan het nadeel hebben dat verouderde profielen niet behandeld worden en de betrouwbaarheid per gegeven daarvan onduidelijk blijft. Echter omdat de enquete waarschijnlijk beter kunnen beantwoorden en het negatief zou zijn wanneer zij zelf kunnen kiezen over welk profiel zij de vragen invullen, is toch gekozen voor het meest recent bewerkte profiel.

7.3.3 Standaardvragenlijsten

Ook is de onderzoeksobjecten gevraagd of zij een standaard vragenlijst hadden. Een standaard vragenlijst is een lijst die op sommige sociale netwerksites wordt aangeboden. Daarnaast zijn er mensen die dergelijke vragenlijsten van een andere site halen en in de vrije ruimte van hun profiel zetten.

De gebruikte vragenlijsten hebben diverse onderwerpen. Op Gurlz.nl staan bijvoorbeeld de vragenlijsten:

- 'Radio, tv, film en muziek'
- 'Favorieten,
- 'Onderwijs, werk en maatschappij'
- 'Liefde en relaties'
- 'Sex'
- 'Uitgaan'
- 'Feesten en evenementen'
- 'Vakantie'
- 'Vrije tijd & sport'
- 'Sex ABC-tje'
- 'About Me (uiterlijk)'

Vragenlijst 'Radio, tv, film en muziek' op Gurlz.nl:

- Welke radiozender heb je meestal aanstaan?
- Van welke soorten muziek houd je?
- Wie is je favoriete radio DJ?
- Wat is televisieprogramma top 3?
- Hoeveel uur per dag hang je voor de buis?
- Leukste man op televisie?
- Leukste vrouw op televisie?
- Leukste tv-commercial aller tijden?
- Bij welke irritante reclame gooi je je televisie echt het raam uit?
- Welke programma moet gelijk van de buis?
- Bij welke televisiepersoonlijkheid krijg je spontaan braakneigingen?
- Beste 3 films aller tijden?
- Lekkerste acteur/actrice?
- Favoriete muziekstijl(en)?

Figuur 7: Voorbeeld vragenlijst

De onderzoeksobjecten is gevraagd of zij een standaardvragenlijst hadden, wat het onderwerp daarvan was en waar de vragenlijst vandaan komt. Bij enquêtes aan de leden van sites die zelf vragenlijsten aanbieden (Hunkz.nl, Gurlz.nl en Ikku.nl), is er expliciet bijgevraagd of de vragenlijst bij die site vandaan kwam. Ook is gevraagd of de antwoorden op de vragenlijst klopten met de werkelijkheid en wat er eventueel niet klopte. De vraag of men een standaardvragenlijst op zijn profiel heeft staan was verplicht, overige vragen over dit onderwerp niet.

7.3.4 Betrouwbaarheid overige informatie

Naast de velden is ook vrije ruimte op het profiel waarop de eigenaar van het profiel kan zetten wat hij wil[1]. De vrije ruimte zal soms ook gevuld worden met de eerder genoemde standaardgegevens of een standaard vragenlijst. Omdat per persoon verschilt hoe de vrije ruimte is ingevuld, is in het algemeen gevraagd of er nog andere informatie op het profiel niet klopt en (wanneer dit het geval is) wat die informatie is. De vraag naar het bestaan van andere foutieve informatie op het profiel is verplicht. De verdiepingsvraag is niet verplicht omdat de site waarop de enquête is afgenomen de mogelijkheid niet bood vragen afhankelijk te maken van het antwoord van de voorgaande vraag.

[1] De invulling van het profiel (dus ook de vrije ruimte) wordt op veel sites beperkt door het gebruikersreglement en de technische mogelijkheden.

7.3.5 Reden foutieve gegevens

Er is gevraagd naar de reden van de foutieve informatie. Dit was geen verplichte vraag, aangezien deze vraag niet essentieel voor het onderzoek is maar enkel meer achtergrondinformatie geeft. Door deze niet verplicht te maken is de kans klein dat deze vraag respondenten afschrikt foutieve gegevens aan te geven.

7.3.6 Persoonsgegevens

Om gegevens uit deze enquête iets te laten zeggen over bepaalde groepen profielen, zijn vragen naar leeftijd, geslacht en opleidingsniveau opgenomen. Deze vragen zijn verplicht gesteld. De vragen zijn opgenomen om de resultaten beter vergelijkbaar te maken met toekomstig onderzoek en mogelijke relaties tussen antwoorden op databetrouwbaarheids- of sociale wenselijkheidsvragen en deze persoonskenmerken te kunnen leggen.

De respondent kan bij het vragen naar persoonsgegevens het vermoeden hebben dat hij aan de hand van de gegevens geïdentificeerd wordt of dat de persoonsgegevens gebruikt worden om het bijbehorende profiel te vinden. Om dit vermoeden zo veel mogelijk te voorkomen wordt maar een klein aantal persoonsgegevens gevraagd. Ook is de respondenten aangegeven dat zij de enquête anoniem kunnen invullen. De personen die kans willen maken op de prijs konden de enquête niet anoniem invullen. Er moet immers contact opgenomen kunnen worden met de prijswinnaar.

7.3.7 Sociale wenselijkheidsvragen

Sociale wenselijkheidsvragen worden gebruikt om in te schatten in hoe verre een persoon sociaal wenselijke antwoorden geeft. Omdat vragen als ‘Klopt de leeftijd die op je profiel staat?’ kunnen worden gelezen als ‘Heb je gelogen?’, kan het zijn dat mensen die foutieve informatie op hun profiel hebben staan, gaan liegen op de vraag of de informatie klopt, omdat eerlijkheid sociaal wenselijk is. Aan de hand van sociale wenselijkheidsvragen wordt dit getest.

Op zoek naar sociale wenselijkheidvragenlijsten, kwam ik erachter dat de vragenlijsten meestal niet aan de artikelen over het onderzoek toegevoegd waren. De vragenlijsten worden wel beschikbaar gesteld aan afgestudeerden en studenten aan psychologische en sociale studies. Daar kon geen gebruik van gemaakt worden aangezien ik niet tot deze groep behoor. Het gevolg hiervan is dat de keuze uit sociale wenselijkheid vragenlijsten die toegepast kunnen worden in dit onderzoek vrij klein is.

De vragenlijst mocht maximaal 15 vragen bevatten, omdat de sociale wenselijkheidsvragen niet de overhand moesten krijgen in de totale enquête. Daarnaast ging de voorkeur uit naar een vragenlijst die was gevalideerd onder leeftijdsgenoten van de onderzoeksobjecten.

Uit het kleine aantal vragenlijsten dat ter beschikking was voor dit onderzoek, werd een selectie gemaakt a.d.h.v. wensen m.b.t. validatiegroepen en hoeveelheid vragen. Deze selectie resulteerde in de keuze voor de ‘verkorte vragenlijst B’ [O32] van de ‘Crandall Social Desirability Test for Children’ (CSDTC). De CSDTC is een sociale wenselijkheidtest die door Crandall et al. ontwikkeld is [O33]. Deze test is door Crandall et al. gevalideerd onder kinderen uit de derde, vierde, vijfde, zesde, achtste, tiende en twaalfde klas. Deze kinderen hebben dus een leeftijd tussen de acht en achttien jaar. De CSDTC is ontwikkeld gepubliceerd in 1965 en wordt nog steeds gebruikt. Versies van de CSDTC zijn recentelijk onder andere gebruikt in onderzoeken gepubliceerd in 2006 [O34] en 2007 [O35] [O36].

Volgens Jurbergs et al. is de betrouwbaarheid en validiteit van dit instrument ruimschoots bewezen [O35]. Hierbij moet wel opgemerkt worden dat niet alle respondenten kinderen zullen zijn en de vragenlijst dus niet op iedere leeftijdscategorie binnen de onderzoeksgroep is gevalideerd. Ook is de vragenlijst in Amerika gevalideerd. Verschillen tussen de validatiegroep en de onderzoeksgroep zijn niet wenselijk, maar kleine verschillen zijn niet te voorkomen. De verschillen tussen validatiegroep en onderzoeksgroep zijn besproken in hoofdstuk 14.

Carifio [O31] heeft diverse methodes getest om de beste korte vragenlijst samen te stellen uit de CSDTC. De twaalf vragen die in dit onderzoek gebruikt worden (‘vragenlijst B’ in Carifio’s artikel [O31]) bleek de grootste intercorrelatie (namelijk 0,92) te hebben met de scores op de CSDTC, in vergelijking met andere verkorte vragenlijsten die hij heeft samengesteld uit de CSDTC. Omdat de totale vragenlijst van de CSDTC te lang was voor dit onderzoek, werd Carifio’s ‘vragenlijst B’ gezien als een goed alternatief.

Short-Form B Items of the Crandall Social Desirability Test for Children
True/ False response format

1. I sometimes feel angry when I don't get my way.
2. I never say anything that would make a person feel bad.
3. Sometimes I argue with my mother to do something she doesn't want me to.
4. I have never been tempted to break a rule or a law.
5. I would never hit anyone who was smaller then I am.
6. Sometimes I do not feel like doing what my teachers want me to do.
7. I am always polite even to people who are not very nice.
8. I never borrow anything without asking permission first.
9. Sometimes I say things just to impress my friends.
10. I am always careful about keeping my cloths neat and my room picked up.
11. Sometimes I don't feel like obeying my parents
12. Sometimes I feel like staying home from school even if I am not sick.

Figuur 8: 'Verkorte vragenlijst B' [O32] van de 'Crandall Social Desirability Test for Children' [O31]

Vragenlijst B bestaat uit 12 vragen en is in het Engels opgesteld. De vragenlijst is door mij van Engels naar Nederlands vertaald, deze vertaling heb ik laten controleren. De Nederlandstalige vertaling van de sociale wenselijkheidsvragen zoals gebruikt in de enquête, staat in figuur 9.

Sociale wenselijkheidsvragen:

Geef aan of de volgende stellingen waar of niet waar zijn

- Ik voel me soms boos als ik mijn zin niet krijg.
- Ik zeg nooit iets dat iemand kan kwetsen.
- Ik ga soms in discussie met mijn ouders als ik iets niet mag doen.
- Ik ben nooit in de verleiding gekomen een regel of wet te breken.
- Ik zou nooit iemand slaan die kleiner is dan ik.
- Ik heb soms geen zin om te doen wat de docent van mij vraagt.
- Ik ben altijd beleefd, zelfs tegen mensen die niet vriendelijk zijn.
- Ik leen nooit iets zonder het eerst te vragen.
- Ik zeg soms dingen puur om indruk te maken op mijn vrienden.
- Ik houd altijd mijn kleding netjes en mijn kamer schoon.
- Ik heb soms geen zin om te doen wat mijn ouders van mij vragen.
- Ik heb soms zin om thuis te blijven van school, zelfs als ik niet ziek ben.

Figuur 9: Vertaalde versie 'Verkorte vragenlijst B'

Binnen dit onderzoek zullen antwoorden van een persoon worden verwijderd wanneer op meer dan 66 procent van de vragen uit de sociale wenselijkheidtest een sociaal wenselijk antwoord wordt gegeven. Dit betekent dat op acht van de twaalf vragen het sociaal wenselijke antwoord wordt gegeven. De inschatting is dat er dan voldoende response behouden blijft om de betrouwbaarheid van de data op profielen te kunnen beoordelen.

De sociale wenselijkheidsvragen zijn in de vragenlijst 'Algemene vragen 2' genoemd, om het doel van deze vragen achter te houden. Als de respondenten het doel van deze vragen voor beantwoording zouden weten, zou dit de antwoorden op de vraag kunnen beïnvloeden. De sociale wenselijkheidsvragen waren verplichte vragen. Deze vragen zijn bewust zo laat mogelijk gesteld. Omdat dit in de ogen van onderzoeksobjecten rare vragen zijn (omdat zij het doel van de vragen niet kennen), kan dit een reden zijn om af te haken. Ik verwacht dat ze minder snel door dit soort vragen zullen afhaken wanneer ze het merendeel van de vragenlijst al ingevuld hebben.

7.3.8 Betrouwbaarheid profielen van vrienden

De onderzoeksobjecten is ook gevraagd of er foutieve informatie in het profiel van hun vrienden staat. Hierdoor kan over meer objecten dan onderzocht zijn bepaald worden waarover personen liegen. Het beantwoorden van deze gegevens zal waarschijnlijk minder betrouwbaar zijn, aangezien die persoon iets zegt over het profiel van een ander. Er vanuit gaande dat hij het profiel van zijn vrienden niet zal opzoeken voor deze enquête, zal hij het profiel van de ander waarschijnlijk minder goed kennen dan het eigen profiel.

Mijn hypothese is dat de onderwerpen waarin foutieve informatie staat, per site verschillen. Daarom zouden foutieve gegevens die op een andere sociale netwerksite staan, niet per definitie iets zeggen over foute gegevens op de onderzochte sociale netwerksites. Dus wordt hen gevraagd op welke site het door hun constateerde oncorrecte profiel staat. Ook werd gevraagd naar de onderwerpen en de reden van de foutieve gegevens. Al deze vragen waren niet verplicht, ook konden onderzoeksobjecten invullen dat ze niet wisten of hun vrienden foutieve informatie ingevuld hadden. Deze vragen geven wel een indruk van onderwerpen die vaak niet kloppen, maar de mate waarin de betrouwbaarheid van het ene onderwerp verschilt van de ander is niet uit de antwoorden op deze vragen af te leiden. Dit maakt vragen over dit onderwerp niet essentieel voor dit onderzoek, waardoor het niet nodig was deze vragen verplicht te stellen.

Het beantwoorden van deze vragen kan worden gezien als het verklikken van vrienden. Dit kan er in resulteren dat men hierover gaat liegen. Om dit te voorkomen hadden de respondenten de keuze de vragen over het onderwerp in zijn geheel niet in te vullen of aan te geven dat men niet weet of vrienden dit doen.

7.3.9 Deelname aan wedstrijd

Als onderzoeksobjecten kans wilden maken op een tijdschriftenbon t.w.v. vijftig euro, moesten zij aan het eind van de vragenlijst hun e-mailadres invullen. Deze vraag is aan het eind gesteld om de kans zo klein mogelijk te maken dat mensen de indruk kregen dat ze verplicht hun e-mailadres in moesten vullen.

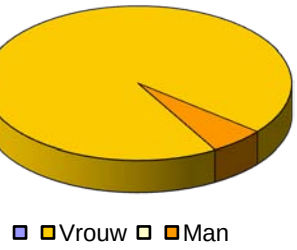
8 Resultaten experiment betrouwbaarheid van data

In dit hoofdstuk zijn de resultaten van het databetrouwbaarheidsonderzoek beschreven. In de volgende paragraven staan de gegeven antwoorden op de vragen uit het vorige hoofdstuk weergegeven. De verwerking van de antwoorden wordt beschreven en de uiteindelijke resultaten van het experiment worden weergegeven.

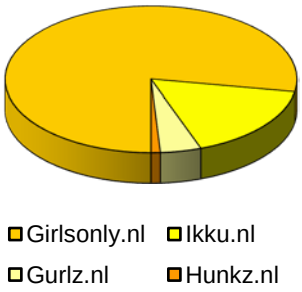
8.1 Respondenten

Uiteindelijk zijn 519 ingevulde vragenlijsten binnengekomen. Het aantal respondenten per site verschilt wel, namelijk 32 voor Hunkz / Gurlz, 89 leden van Ikku en 398 leden van Girlsonly. Ook in geslacht zijn er grote verschillen. Dit komt onder andere door de grote deelname van Girlsonly leden. Meer dan de helft van de respondenten heeft een opleidingsniveau behorende bij de middelbare school (deels een gevolg van de leeftijd van de respondenten). In onderstaande figuren staan deze en andere verschillen tussen respondenten op verschillende eigenschappen weergegeven. Hiermee wordt getracht een indruk te geven van de respondenten. Uit de verschillen kan afgeleid worden dat de resultaten een vertekend beeld kunnen geven, omdat een aantal eigenschappen een onverwachte verdeling vormen. Een voorbeeld van een eigenschap met een onverwachte verdeling is het lidmaatschap van de sociale netwerksite. Ikku heeft bijvoorbeeld meer leden dan Girlsonly, terwijl meer Girlsonly leden de vragenlijst hebben beantwoord, dit kan een vertekend beeld geven.

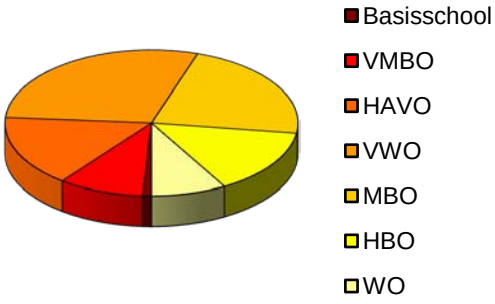
Figuur 10: Geslacht respondenten

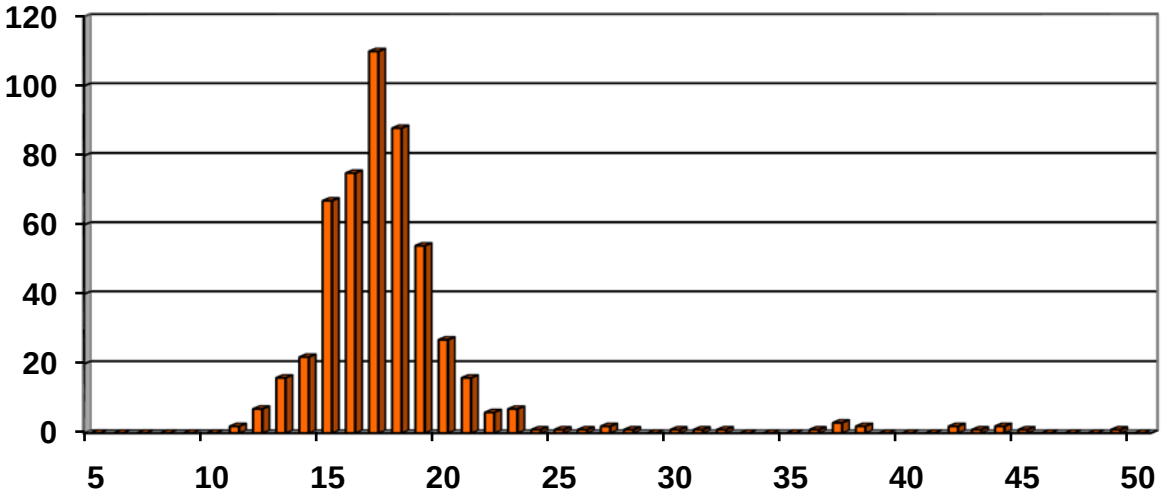


Figuur 11: sociale netwerksite van respondenten



Figuur 12: Opleidingsniveau respondenten





Figuur 13: Leeftijd respondenten

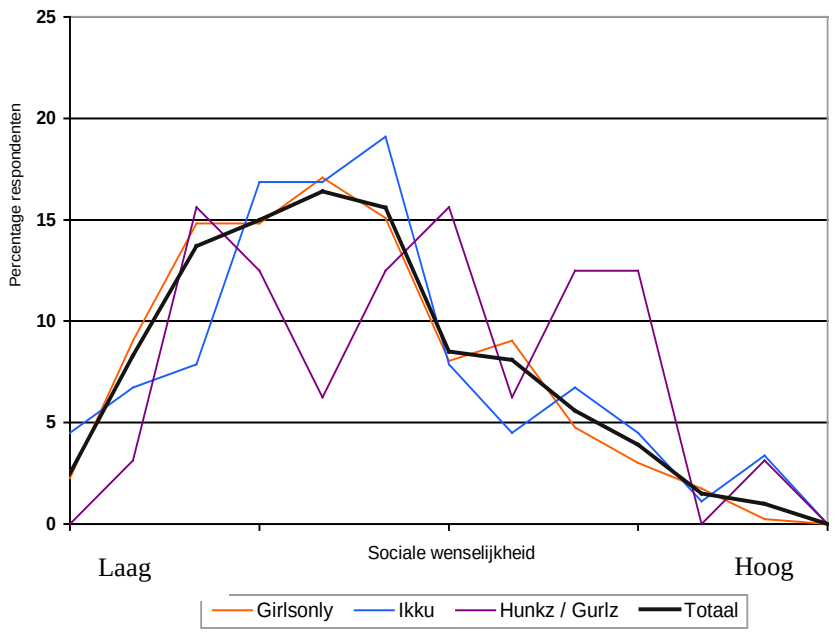
Site	Gemiddeld	Mediaan	Standaard deviatie
Ikku.nl	19	16	8,6
Girlsonly.nl	17	17	1,9
Gurlz.nl & Hunkz.nl	21	18	8,1

Tabel 4: Leeftijd van respondenten

8.2 Sociale wenselijkheid

In paragraaf 7.3.7 is uitgelegd wat sociale wenselijkheid is en hoe de sociale wenselijkheid binnen dit experiment gemeten wordt. De sociale wenselijkheid van de verschillende sociale netwerksites staat in Tabel 5 en Figuur 14 weergegeven. Er zitten een aantal grote verschillen tussen de verschillende sociale netwerksites. Deze verschillen zullen onder andere veroorzaakt worden door het grote verschil in aantal respondenten. Wanneer er maar een klein aantal respondenten bij een site horen, heeft de score van één persoon een groot effect op de loop van de lijn in onderstaand diagram.

Het gedeelte van de vragenlijst dat de sociale wenselijkheid testte, telt twaalf vragen. Op elk van deze vragen kan een persoon al dan niet het sociaal wenselijke antwoord geven. Wanneer een persoon het sociaal wenselijke antwoord op een vraag heeft gegeven wordt zijn score met één opgehoogd. Op deze manier wordt de sociale wenselijkheid van de persoon berekend aan de hand van de twaalf vragen. Een persoon kan een score hebben van nul tot en met twaalf. Nul is de score die aangeeft dat iemand helemaal geen sociaal wenselijke antwoorden heeft gegeven, deze persoon is dus niet snel geneigd zijn sociaal wenselijke antwoorden te geven op de overige vragen. Een persoon met de score twaalf heeft alle vragen sociaal wenselijk beantwoord en is dus waarschijnlijk sterk geneigd sociaal wenselijke antwoorden te geven op de overige vragen. De vraagstelling is dusdanig dat het zeer onwaarschijnlijk is dat iemand daadwerkelijk het sociaal wenselijke gedrag vertoont op alle punten waar naar gevraagd wordt.



Figuur 14: Sociale wenselijkheid per sociale netwerksite en in totaal

Sociale wenselijkheid	0	1	2	3	4	5	6	7	8	9	10	11	12
Girlsonly	2,3	9,1	14,8	14,8	17,1	15,1	8,1	9,1	4,8	3,0	1,8	0,3	0,0
Ikku	4,5	6,7	7,9	16,9	16,9	19,1	7,9	4,5	6,7	4,5	1,1	3,4	0,0
Hunkz/Gurlz	0,0	3,1	15,6	12,5	6,3	12,5	15,6	6,3	12,5	12,5	0,0	3,1	0,0
Totaal aantal respondenten 519	13	43	71	78	85	81	44	42	29	20	8	5	0
Totaal percentage	2,5	8,3	13,7	15,0	16,4	15,6	8,5	8,1	5,6	3,9	1,5	1,0	0,0

Tabel 5: Percentages sociale wenselijkheidsscore per sociale netwerksite

In tabel 5 en figuur 14 kan de sociale wenselijkheid van alle respondenten gezamenlijk afgelezen worden. Zoals in hoofdstuk 5 gelezen kan worden geeft de sociale wenselijkheidsscore aan in hoe verre iemand geneigd is sociaal wenselijke antwoorden te geven. Het onjuist weergeven van informatie zal vaak als sociaal onwenselijk beschouwd worden, zeker wanneer er sprake is van een bewuste leugen. Daarom wordt verwacht dat personen die geneigd zijn eerder een sociaal wenselijk antwoord te geven, minder vaak aangeven foutieve informatie op hun profiel te hebben. De betrouwbaarheid van de antwoorden van een persoon is waarschijnlijk groter wanneer iemand een lage sociale wenselijkheidsscore heeft.

In paragraaf 7.3.7 is al eerder aangegeven dat de antwoorden van de mensen met een hoge sociale wenselijkheidsscore verwijderd zouden worden. Aan de hand van tabel 5 en figuur 14 is er voor gekozen de enquêtes tot en met een sociale wenselijkheidsscore van acht te behouden en de enquêtes met een sociale wenselijkheidsscore van 9 of hoger te verwijderen. Dit betekent dat 93,7 procent van de enquêtes behouden blijft voor het onderzoek. De verwijderde 6,3 procent van het totaal, zijn 33 responses. Van de response afkomstig van Girlsonly is vijf procent verwijderd, van de response van Ikku.nl is negen procent verwijderd en van de responses van Hunkz en Gurlz is 21 procent verwijderd. Dit is af te lezen uit tabel 3.

Bij response van Hunkz.nl is 44 procent verwijderd, terwijl slechts 4 procent van de response van Gurlz verwijderd is. Dit geeft de indruk dat mannen in deze enquête een hogere sociale wenselijkheidsscore hebben, maar de cijfers van Hunkz en Gurlz geven een ander beeld dan de andere sites. Er zijn in totaal 6 mannelijke respondenten verwijderd uit een totaal van 30 (20 procent). Hiervan zijn er 4 afkomstig van Hunkz. Bij Hunkz.nl is 44 procent van het totaal aantal mannelijke respondenten verwijderd, terwijl dit bij Ikku.nl op 11 procent ligt.

8.3 Aantal profielen

Aantal profielen	Percentage respondenten
0	0,6
1	90,1
2	8,0
3	1,0
4	0,2

Tabel 6: Aantal profielen van respondenten met sociale wenselijkheid < 9

Negentig procent van de respondenten heeft één profiel op de betreffende sociale netwerksite. Ongeveer acht procent geeft aan twee profielen te hebben. De percentages zijn terug te vinden in Tabel 5. Het verschil tussen de mensen met een hoge en lage sociale wenselijkheid is niet significant.

Er is één persoon (met sociale wenselijkheidsscore 9) die aangeeft 8861 profielen te hebben op de betreffende site. Ik verwacht dat de persoon een absurd hoog getal heeft ingevoerd omdat de persoon het echte antwoord niet heeft willen geven. Dit lijkt op het gedrag dat Boyd ontdekte op sociale netwerksites. Uit haar onderzoek bleek dat sommige mensen wanneer zij liegen over hun postcode de voorkeur hebben een postcode in te voeren waarvan zij dachten dat die niet kon bestaan, als 1234 AB [O2]. De persoon kan ook een fout hebben gemaakt. De kans lijkt me erg klein dat hij het werkelijke aantal heeft opgegeven.

De kans dat deze persoon de waarheid heeft gesproken over zijn aantal profielen lijkt mij erg klein. Deze persoon geeft aan verder geen foutieve data op haar profiel te hebben staan en heeft geen reden aangegeven voor het hebben van meerdere profielen. Uit deze gegevens zou een inschatting gemaakt kunnen worden hoe waarschijnlijk het bezit van 8861 profielen door die persoon is. Wanneer deze persoon een reden aan had gegeven, kan aan de hand van de reden deze bewering eventueel inschatten als meer betrouwbaar.

8.3.1 Redenen voor meerdere profielen

Vierenveertig respondenten hebben een reden aangegeven voor het bezit van meerdere profielen. Bij deze respondenten zitten ook personen met een hoge sociale wenselijkheid.

Reden	respon-denten
Anonimiteit	39 procent
Geen toegang oud profiel	36 procent
Naam / Nickname wijzigen	5 procent
Privacy	5 procent
Anders	20 procent

Tabel 7: Redenen van meerdere profielen ongeacht sociale wenselijkheid

Negenendertig procent van de respondenten die een reden hebben aangegeven, zeggen dat anonimiteit voor hen de reden is voor het bezitten van meerdere profielen. Het grootste deel van de respondenten die dit antwoord gaven, zegt één profiel te hebben waar men de identiteit kenbaar maakt en een ander profiel om anonieme berichten mee te kunnen posten.

Zesendertig procent van de respondenten die een reden heeft aan gegeven, zegt een nieuw profiel te hebben aangemaakt omdat zij geen toegang meer tot het oude profiel hadden. De meeste respondenten die dit antwoord hebben gegeven zijn de inlognaam of het wachtwoord van de sociale netwerksite vergeten of hebben geen toegang meer tot het e-mailadres waar hun profiel aan gekoppeld is. Ook waren er respondenten die zeiden geen toegang meer te hebben tot hun profiel als gevolg van een storing op de site.

Vijf procent gaf aan dat zij hun naam of nickname wilden veranderen, maar dat dit op de betreffende site niet mogelijk was binnen hetzelfde profiel. Om toch onder een andere naam verder te kunnen gaan hebben deze mensen een nieuw profiel aangemaakt.

Andere redenen zijn per reden door één respondent (2 procent van de personen die een reden heeft opgegeven) genoemd. Andere redenen waren:

- Het account is bewust geblokkeerd door de beheerders van de site. Om toch nog op de site te kunnen werd een nieuw account aangemaakt.
- Vergeten dat er al een profiel in het bezit was.
- Er waren veel foutmeldingen op het profiel. Om uit te testen of dit aan dit profiel lag werd een nieuwe aangemaakt.
- Er werd er een voor zichzelf gemaakt en één voor een vriend.
- Het tweede profiel dient als reserve. Als reden wordt aangegeven dat een profiel het soms niet doet.
- Een nieuw profiel werd aangemaakt omdat de persoon informatie wilde wijzigen, maar geen zin had het oude profiel leeg te maken alvorens deze te vullen met nieuwe informatie.
- “Ik weet het niet.”
- Men vond het leuk meerdere profielen te bezitten. Wat hier de reden voor is wordt niet aangegeven.
- Uitgekeken op het oude profiel.

8.3.2 Verschillen tussen de profielen

Aan de respondenten werd gevraagd wat het verschil was tussen de profielen. Dit was een niet verplichte vraag, die is beantwoord door 21 personen. Door 52 procent van de respondenten die een verschil hebben aangegeven, wordt opgegeven dat het ene profiel anoniem is en het andere niet. Het anonieme profiel zal waarschijnlijk zo min mogelijk informatie bevatten. Waar men verplicht gegevens in dient te vullen, verwacht ik dat zij, wanneer het te persoonlijk wordt, zullen liegen. Achttien procent van de respondenten die anonimiteit als reden aangegeven hebben voor het verschil tussen de profielen die zij bezitten, geven expliciet aan dat zij bewust foutieve informatie geplaatst hebben om deze anonimiteit te waarborgen. Van de overige respondenten met meerdere profielen, waaronder een anoniem profiel, is niet bekend of zij foutieve informatie opgeven om hun identiteit te waarborgen.

Een ander genoemd verschil tussen de profielen is dat de een actuele informatie bevat en de ander verouderde informatie. Dit wordt genoemd door 38 procent van de respondenten die een verschil hebben aangegeven. Het is afhankelijk van de persoon welke data verschilt, omdat dit ligt aan de veranderingen in zijn eigenschappen. De veranderde eigenschappen zullen in dit geval waarschijnlijk geen geboortedatum, geslacht, seksuele voorkeur, kleur ogen of huidskleur zijn, maar eerder hobby's, favoriete artiest, woonplaats of lievelingseten.

Een tweetal respondenten heeft aangegeven wat het precieze verschil is tussen de meerdere profielen. De eerste geeft aan dat de naam het verschil is tussen de profielen; de reden van het aanmaken van een nieuw profiel was voor deze persoon was dat hij niet meer kon inloggen op zijn oude profiel. Een ander die om dezelfde reden een nieuw profiel aanmaakte, geeft aan dat op het ene profiel persoonlijke informatie staat en enkel tekst staat en op het andere alleen foto's staan. Deze personen representeren afzonderlijk elk ongeveer vijf procent van het totaal aantal respondenten dat deze vraag heeft beantwoord.

8.4 Aantal fouten

In Tabel 8 zijn het aantal respondenten (met een sociale wenselijkheid < 9) weergegeven met het aantal fouten dat men aangaf in het profiel te hebben op de gevraagde twaalf onderwerpen. De respondenten zijn gegroepeerd op het aantal onderwerpen dat foutief in het profiel was ingevuld. De twaalf onderwerpen waar hier over gesproken wordt, zijn de onderwerpen die terug te vinden zijn in tabel 6. Van deze twaalf onderwerpen is specifiek gevraagd of de informatie betreffende het onderwerp klopt of niet. Deze vragen waren verplicht. In Tabel 9 zijn de aantallen respondenten uit tabel 8 opgehoogd omdat respondenten in open vragen hebben aangegeven dat zij nog op andere punten foutieve informatie op hun profiel hebben staan.

Aantal fouten per profiel	Aantal respondenten
0	407
1	46
2	18
3	5
4	5
5	1
7	3
10	1
Aantal profielen met fouten	79
Percentage van totaal	16,3 procent

Tabel 8: Respondenten met fouten op 12 onderwerpen in het profiel - Uit respondenten met een sociale wenselijkheid <9

Als wij deze aantallen vergelijken met de cijfers in paragraaf 6.1 zien we dat het percentage fouten hoger ligt dan het aantal leugens in e-mail, maar lager dan het aantal leugens in andere media. Hier mag echter niet uit geconcludeerd worden dat er op sociale netwerksites meer gelogen word dan in e-mail, wegens de hieronder beschreven redenen.

Ten eerste is er in de enquête gevraagd naar het niet kloppen van informatie, in plaats van naar leugens. Het aantal leugens op sociale netwerksites zal waarschijnlijk lager liggen dan de percentages weergegeven in tabel 8.

Aantal fouten per profiel	Aantal respondenten
0	392
1	58
2	17
3	8
4	5
5	1
6	1
7	3
10	1
Aantal profielen met fouten	93
Percentage van totaal	19,1 procent

Tabel 9: Respondenten met fouten in het profiel op alle onderwerpen - Uit respondenten met een sociale wenselijkheid <9

Ten tweede is er binnen dit onderzoek een andere manier van rapporteren aangehouden. Men is niet gevraagd een bepaalde tijd al zijn communicatie bij te houden en de leugens daarbinnen. Er is gevraagd achteraf te noemen hoeveel fouten er zitten in de informatie. Men registreert niet actief de fouten maar doet dit achteraf, ook kan er niet vanuit worden gegaan dat respondenten hun profiel opzoeken bij het invullen van de enquête om het profiel te controleren op fouten. Hierdoor kan men fouten in het profiel vergeten zijn. Mensen stellen een profiel meestal niet in een keer samen, terwijl dat bij een e-mail en de andere in [O21] genoemde media meestal wel het geval is. De eventuele leugens in een profiel kunnen in meerdere gebruikerssessies ontstaan zijn.

Dat het percentage in de buurt ligt van het percentage leugens in andere media, verteld ons wel dat waarschijnlijk een groot deel van de leugens boven water is gekomen.

8.5 Foutieve data

Zoals in hoofdstuk 5 is beschreven, werd verwacht dat mensen met een hogere sociale wenselijkheid minder vaak aangeven foutieve informatie op hun profiel te hebben, omdat zij dit kunnen zien als het toegeven van een leugen.

	Leeftijd	Geslacht	Prov inc i e	Woonplaats	Opleiding	Relatiestatus	Geaardheid	Huidskleur	Oogkleur	Ui ter l ijk	Hobby's	Sterrenbeeld
Klopt	94,0	98,4	85,4	68,9	65,0	61,9	64,2	58,2	60,7	56,4	71,0	84,4
Klopt niet	3,1	0,6	2,1	7,0	4,1	3,9	2,1	1,4	1,6	1,6	1,9	1,4
Niet ingevuld	2,9	1,0	12,6	24,1	30,9	34,2	33,7	40,3	37,7	42,0	27,2	14,2

Tabel 10: Percentage kloppende gegevens van respondenten met sociale wenselijkheidsscore < 9

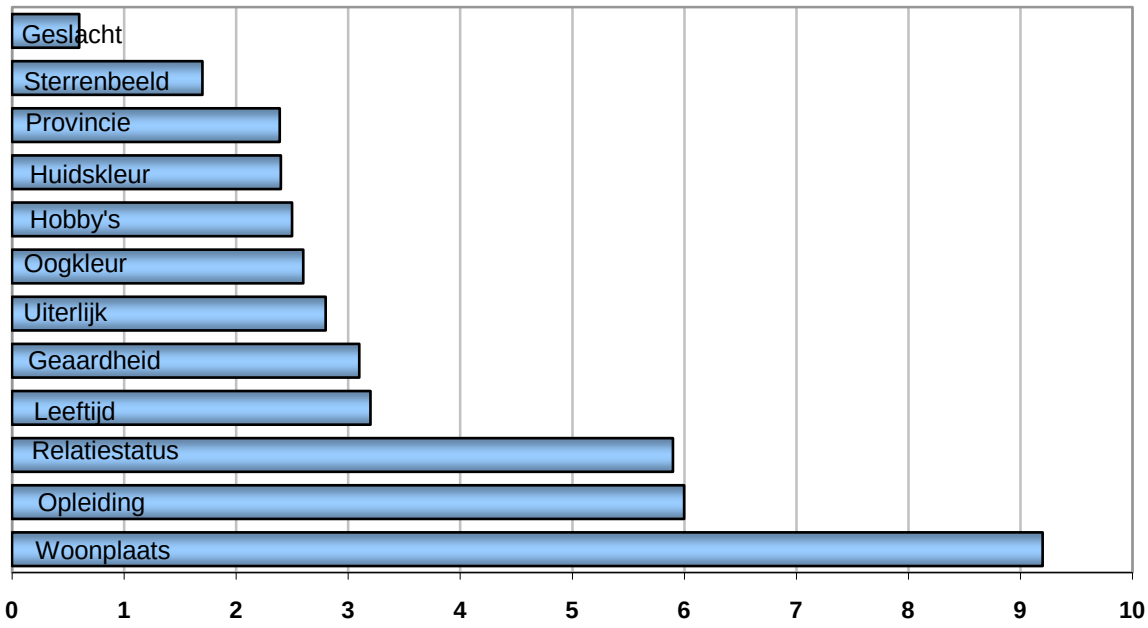
Uit een vergelijking tussen de cijfers van mensen met een lage en een hoge sociale wenselijkheid, blijkt dat de mensen met een hoge sociale wenselijkheid op het grootste gedeelte van de onderwerpen aangeven minder of evenveel foutieve data in hun profiel hebben staan als mensen met een hoge sociale wenselijkheid. Deze verschillen zijn echter niet significant.

Om te berekenen hoeveel van de ingevulde data foutief is, zijn de absolute getallen behorende bij tabel 10 gebruikt, deze zijn terug te vinden in de bijlage. Per onderwerp is het aantal dat bij ‘klopt’ staat opgeteld bij het aantal dat bij ‘klopt niet’ staat. Dit is dus het aantal respondenten dat over dat onderwerp iets ingevuld heeft op zijn profiel en een sociale wenselijkheidsscore lager dan negen heeft. In tabel 11 zijn de percentages voor de wel en niet kloppende data van de totaal ingevulde data voor dat onderwerp weergegeven. Verschillen tussen de percentages uit tabel 10 en 11 worden veroorzaakt door het buiten beschouwing laten van de respondenten die aangegeven dat zij niet iets ingevuld hebben over dat onderwerp. Hierbij dient de kanttekening te worden gemaakt dat het aantal respondenten dat aangaf de gegevens niet te hebben ingevuld, per onderwerp verschillen.

	Leeftijd	Geslacht	Prov inc i e	Woonplaats	Opleiding	Relatiestatus	Geaardheid	Huidskleur	Oogkleur	Ui ter l ijk	Hobby's	Sterrenbeeld
Klopt	96,8	99,4	97,6	90,8	94	94,1	96,9	97,6	97,4	97,2	97,5	98,3
Klopt niet	3,2	0,6	2,4	9,2	6	5,9	3,1	2,4	2,6	2,8	2,5	1,7

Tabel 11: Percentage kloppende informatie van ingevulde onderwerpen.

Figuur 16 is een Top 12, opgesteld is aan de hand van gegevens uit Tabel 11. Wanneer de top 12 opgesteld zou zijn aan de hand van de gegevens in Tabel 10, zou dit een vertekend beeld geven. Dit komt doordat er minder vaak gelogen kan worden over gegevens die minder ingevuld worden, Minder ingevulde gegevens zouden dan automatisch lager in de top 12 komen, ook wanneer er relatief gezien evenveel over gelogen wordt.



Figuur 15: Top 12 Foutieve informatie op profielen wanneer dit onderwerp ingevuld is, gebaseerd op Tabel 6.

1. Woonplaats
2. Opleiding
3. Relatiestatus
4. Leeftijd
5. Geaardheid
6. Uiterlijk
7. Oogkleur
8. Hobby's
9. Huidskleur
10. Provincie
11. Sterrenbeeld
12. Geslacht

Figuur 16: Top 12 Foutieve informatie

Onderwerpen die aangegeven zijn in de open vraag als niet kloppende gegevens zijn:

- Naam
- Achternaam
- Geboortedatum (jaartal klopt wel)
- Idolen
- Favoriete films
- Favoriete muziek
- Huisdieren (hond is overleden)

Naar de namen en geboortedatum heb ik bewust niet gevraagd, omdat deze niet van belang waren. Er is echter wel gevraagd of de leeftijd klopt. Favoriete films, muziek en huisdieren behoren niet tot de meest ingevulde velden op sociale netwerksites, vandaar dat daar niet naar gevraagd is. Er is echter wel naar onderwerpen gevraagd die normaal niet tot de meest ingevulde behoren op sociale netwerksites, namelijk huidskleur en kleur ogen. Omdat deze onderwerpen op een van de medewerkende sites (Ikku.nl) verplicht ingevuld moesten worden bij het inschrijven en dus veel ingevuld is bij die site, zijn deze onderwerpen wel opgenomen in de vragenlijst.

8.6 Standaard vragenlijst

Mensen hebben soms een standaard vragenlijst op hun profiel staan. Wat er wordt verstaan onder een dergelijke vragenlijst is uiteengezet in paragraaf 7.3.3.

Site	Totaal aantal	Aantal met vragenlijst	Aantal met vragenlijst van totaal
Girlsonly	378	12	3,2 procent
Ikku.nl	89	46	51,7 procent
Gurlz.nl	22	9	40,9 procent
Hunkz.nl	5	1	20,0 procent
Totaal	494	68	13,8 procent

Tabel 12: Aantal respondenten met vragenlijst per site

Het aantal respondenten dat aangeeft een vragenlijst op zijn profiel te hebben staan, is ongeveer veertien procent. De verschillen per site in het plaatsen van een vragenlijst op het profiel zijn opvallend. Deze verschillen zijn uit tabel 12 af te lezen. Op Ikku.nl blijkt het bezit van een vragenlijst populairder te zijn dan op bijvoorbeeld Girlsonly. De kanttekening die hierbij geplaatst dient te worden is dat op Ikku.nl, Hunkz.nl en Gurlz.nl vragenlijsten aangeboden worden.

Van alle respondenten heeft er slechts één aangegeven te hebben gelogen op de vragenlijst die op zijn profiel staat. Wel geeft een ander aan niet te hebben gelogen, maar op sommige vragen niet serieus antwoord te hebben gegeven. Zij heeft geprobeerd een grappige invulling aan een aantal van de vragen te geven.

9 Genereren van associatieve regels

Nu de databetrouwbaarheid onderzocht is, kan gestart worden met het datamining experiment op de data van profielen. In dit hoofdstuk wordt de keuze gemaakt voor een datamining algoritme. Vervolgens wordt uitgelegd op welke manieren associatieve regels gegenereerd kunnen woorden. Tenslotte wordt onderscheid gemaakt tussen verschillende algoritmes om de keuze van het algoritme te onderbouwen.

Om dat associatie analyse een vorm van datamining is, wordt datamining en het gebruik ervan in combinatie met internet eerst kort toegelicht.

9.1 Datamining

De term datamining wordt gebruikt voor een breed scala aan activiteiten welke het doel hebben nieuwe informatie te ontdekken uit een originele dataset. Meestal is de originele dataset niet verzameld voor datamining, maar voor een totaal ander doel [O37]. Datamining taken kunnen op verschillende manieren worden gecategoriseerd, bijvoorbeeld naar doel, input of output.

Wanneer we naar het doel kijken, kunnen we datamining taken indelen in twee categorieën: voorspellend en beschrijvend. In de voorspellende categorie wordt de waarde van een bepaald attribuut ingeschat op basis van andere attributen. De beschrijvende datamining taken proberen patronen (correlaties, trends, clusters, e.d.) te ontdekken die de onderliggende relaties in de data samenvatten [O38]. Associatie analyse, clustering en uitbijter detectie zijn beschrijvende datamining taken.

Er zijn twee types voorspellende datamining taken: classificatie en regressie. Het bepalen of een aanvrager van een lening voldoende kredietwaardig is, is een classificatietaak. Wanneer je de toekomstige prijs van een aandeel wilt voorspellen is dit een regressietaak omdat de prijs van een aandeel een continu attribuut is.

Een ander punt waarop algoritmen gecategoriseerd kunnen worden is het al dan niet bekend zijn van de output. Supervised learning algoritmes karakteriseren zich doordat de output vooraf al bekend is. Een voorbeeld hiervan zijn classificatie-algoritmes. Voor het classificeren zijn de klassen al bekend, het is echter voor uitvoering onbekend wat de regels zijn die er voor zorgen dat een object in een klasse terechtkomt en wat de verdeling van de klassen is. Bij unsupervised learning algoritmes is er vooraf nog geen output bekend. Clustering is hier een voorbeeld van. Wanneer output vooraf nog niet bekend is wordt ook wel gezegd dat er ‘no a priori output’ is.

Daarnaast wordt er ook wel eens gecategoriseerd op herkomst van de data, de categorieën die men kan gebruiken liggen hierbij minder vast, maar men kan denken aan ‘webmining’, ‘boodschappenmandjesanalyse’, e.d.

9.1.1 Datamining gerelateerd aan internet

De term 'Webmining' (datamining gerelateerd aan het internet) werd volgens Kasala en Blockeel [O39] als eerste door Etzioni [O40] gebruikt in 1996. Hij definieerde 'Webmining' als het gebruik van datamining technieken met als doel automatisch informatie te ontdekken en vergaren uit webdocumenten en services. Omdat dit onderzoek valt onder webmining, verdiepen we ons verder in dit onderwerp.

Webmining wordt door Kosala en Blockeel [O39], Madria et al [O41] en Borgers en Levene [O42] ingedeeld in 3 categorieën:

- Web content mining
  Web content mining beschrijft datamining van informatie die geplaatst is op het internet, bijvoorbeeld toegepast door Adomavicius en Tuzhilin [O43]. Web content mining is het werkgebied van dit datamining experiment. Daarom wordt deze vorm van webmining verderop meer uitgelicht.
- Web structure mining
  Web structure mining probeert de onderliggende linkstructuren van het internet te ontdekken [O39]. Web structure mining is o.a. toegepast door Mika [O8].
- Web usage mining
  Web usage mining focust op technieken die het gedrag van gebruikers op het internet probeert te voorspellen. De data die gebruikt wordt is secondaire data op het web als resultaat van interacties, bijvoorbeeld logdata van servergebruik [O39].

Het experiment zoals beschreven in hoofdstuk 9 en 10 valt in de categorie content mining, aangezien het data die op het internet staat gebruikt wordt en niet gefocust wordt op de linkstructuren en het gedrag van gebruikers op het internet niet probeert te voorspellen. Dit experiment probeert wel te raden naar eigenschappen en gedrag, maar het niet te voorspellen. Daarnaast vindt het grootste deel van de eigenschappen en het gedrag dat geraden wordt, offline plaats.

Kasala en Blockeel [O39] geven aan dat web content mining het dataminen van tekst, links en zelfs profielen (die gegenereerd zijn door de site of aangemaakt zijn door de gebruiker) omvat. Profielen die gebruikt worden voor datamining bevatten bijvoorbeeld informatie over aankopen of kenmerken van het surfgedrag op een bepaalde site [O43].

9.2 Eerder wetenschappelijk onderzoek

De meeste wetenschappelijke onderzoeken naar datamining focussen zich op een van de volgende onderwerpen:

- Introductie van nieuwe datamining algoritmes
- Verbetering van bestaande datamining algoritmes
- (nieuwe) Toepassingen van datamining
- Combinatie van datamining technieken en technieken uit aangrenzende vakgebieden (bijvoorbeeld Information Retrieval)

Dit onderzoek focust op databetrouwbaarheid en datamining van profielen om aan de hand daarvan er achter te komen of datamining een goede optie zou zijn om gericht te adverteren op sociale netwerksites. Dit onderzoek past dus in de categorie '(nieuwe) toepassingen van datamining'.

9.2.1 Datamining op persoonlijke data

Artikelen waarin datamining wordt toegepast op profielen met persoonlijke eigenschappen van mensen, met associatieve of causale regels als resultaat, heb ik binnen de onderzoeksperiode nauwelijks kunnen vinden. Daarom is mijn aanname dat associatie analyse weinig op die manier wordt toegepast. Dit onderzoek zal een bijdrage leveren op de vraag of associatie analyse met goede resultaten kan worden toegepast op profielen van mensen. Een tweetal artikelen heb ik kunnen vinden die associatie analyse toepassen op persoonlijke informatie; namelijk een onderzoek naar geweldplegers [O44] en een onderzoek naar klanten van een bank [O45].

9.2.2 Content mining op sociale netwerksites

Hoewel Kasala en Blockeel [O39] aangeven dat web content mining het dataminen van tekst, links en zelfs profielen (die gegenereerd zijn door de site of aangemaakt zijn door de gebruiker) omvat, heb ik nog geen voorbeeld vinden van het toepassen van content mining op sociale netwerksites.

Het onderzoek van Lui en Maes [O27], dat machine learning toepast op de content van sociale netwerksites, komt echter dicht in de buurt van content mining. Dat onderzoek [O27] gebruikt de inhoud van één veld (passions) op de profielen van twee sociale netwerksites om een interestmap aan te maken. Daarmee kunnen relaties tussen eigenschappen en identiteiten in worden gezien. De identiteiten die zij gebruiken hebben zij zelf aangemaakt aan de hand van waardes in het veld 'passions'. Een identiteit kan bijvoorbeeld 'book lover' zijn, samengesteld aan de hand van woorden als: boeken, literatuur, romans en lezen. Zij gebruiken verder geen andere informatie uit het profiel. Verschillen met het experiment beschreven in hoofdstuk 10 zijn:

- het gebruik van machine learning vs. datamining
- het gebruik van het eigenschappen uit één veld vs. de eigenschappen uit het gehele profiel (excl. connecties).
- het aanmaken van identiteiten aan de hand van eigenschappen vs. het groeperen van woorden (voor groeperen van woorden zie paragraaf 10.4.5).

Een nadeel van het gebruik van alle velden van een profiel, maakt dat de informatie uit de veldnaam verloren kan gaan. De veldnaam kan bijvoorbeeld een indruk geven of iemand positief of negatief tegenover de inhoud van het veld staat. Omdat er in het onderzoek van Lui en Maes [O27] maar voor één veld is gekozen, is daar duidelijk dat de eigenaar van het profiel, positief denkt over de ingevoerde woorden. Gebruik van alle informatie op het profiel (inclusief de losse tekst) heeft als voordeel dat meer specifieke karakteristieken van mensen ook naar voren komen, bijvoorbeeld favoriete merken. Daarnaast kunnen onderwerpen waar iemand een negatieve mening over heeft ook erg waardevol zijn.

Lui en Maes [O27] ontwikkelen identiteiten aan de hand van eigenschappen, dit generaliseert gebruikers. Het groeperen van woorden - dat binnen dit onderzoek gebeurt - is minder vergaand, de groepen zijn klein gehouden om geen specifieke eigenschappen te verliezen. In sommige gevallen is naast een aantal kleine groepen ook een grote groep gemaakt waar woorden uit de kleine groepen in voorkomen. Een voorbeeld is de woordgroep ‘eten’ waar o.a. de woordgroepen ‘fastfood’, ‘pasta’, ‘restaurant’ en ‘maaltijd’ in voorkomen.

9.2.3 Datamining op sociale netwerken

Onderzoek naar datamining van profielen op sociale netwerksites heb ik tijdens de onderzoeksperiode niet kunnen vinden. Finin et al. [O46] geven wel aan dat er mogelijkheden zijn voor het toepassen van datamining op sociale netwerksites. Dit experiment kan een van de eerste wetenschappelijke experimenten zijn dat datamining toe te past op profielen van sociale netwerksites. Het is mogelijk dat grote sociale netwerksites datamining al toepassen op de profielen maar dit niet in de publiciteit brengen, bijvoorbeeld uit concurrentieoverwegingen.

Datamining wordt al wel toegepast op sociale netwerken, maar onderzoek naar datamining op de sociale netwerksites die de netwerken faciliteren heb ik niet kunnen vinden. Singh et al. [O47], Mika [O8] en Golder et al. [O48] hebben onderzoek gedaan dat in de buurt komt van datamining op sociale netwerksites, bijvoorbeeld datamining op offline sociale netwerken.

Golder et al. [O48] hebben de communicatie tussen gebruikers van Facebook geanalyseerd. Hiervoor hebben zij statische methodes gebruikt. Dezelfde inputdata hadden zij ook kunnen analyseren door middel van web usage mining, het gebruik van dergelijke dataminingtechnieken had wel tot een ander soort output geleid.

Omdat het idee achter web structure mining afkomstig is uit studies naar offline sociale netwerken, is het logisch dat deze manier van datamining ook toegepast kan worden op online sociale netwerken, een voorbeeld hiervan heb ik echter niet binnen de onderzoeksperiode kunnen vinden. Iets dat hier wel in de buurt komt is het gebruik van visualisatie-tools [O49], waarmee het online sociale netwerk en daarmee de structuur van bijvoorbeeld vriendengroepen inzichtelijk kan worden gemaakt.

9.6 Moeilijkheden bij het toepassen van datamining op profielen van sociale netwerksites

Bij het toepassen van datamining op profielen van sociale netwerksites kunnen diverse moeilijkheden verwacht worden, met betrekking tot de input. Profielen op sociale netwerksites kennen namelijk specifieke eigenschappen, welke datamining bemoeilijken, bijvoorbeeld taalgebruik, gebruik van homoniemen en de manier waarop formulieren op sociale netwerksites zijn opgebouwd.

9.6.1 Taalgebruik

Taalgebruik wordt gebruikt als indicatie van groepsidentiteit [O4]. Nieuwe woorden worden verzonnen en normale woorden krijgen soms een nieuwe betekenis, voorbeelden hiervan zijn: spam, trol en newbie. Volgens Donath [O4] drukt iemand met het gebruik van dergelijke woorden zijn verbondenheid met de online community uit, het is als verhuizen naar een andere regio en het locale accent van de nieuwe regio gaan gebruiken. Om bij de groep te horen hebben veel tieners die zichzelf ander taalgebruik aangeleerd, bijvoorbeeld ‘kewl’ voor ‘cool’ of het willekeurig hoofdletter gebruik in woorden [O4]. Ook worden letters vervangen door cijfers.

Een voorbeeld van taalgebruik die gebruikt wordt als indicatie van groepsidentiteit is breezertaal, een vorm van jongerentaal. Dit taalgebruik komt veel voor op sociale netwerksites. In breezertaal worden hoofdletters en kleine letters afgewisseld, klanken geschreven zoals je ze uitspreekt en cijfers dienen vaak als letters [M42]. Volgens het Van Dale woordenboek is de definitie van breezertaal: vooral door jongeren gehanteerde, sterk fonetische spelling waarin cijfers en andere tekens staan voor de klank van hun uitspraak in het Engels en waarin hoofdletters en kleine letters elkaar regelmatig afwisselen.

Cijfers worden binnen breezertaal ook wel gebruikt op andere manieren dan hetgeen het Van Dale woordenboek noemt. Bijvoorbeeld het uiterlijk van een cijfer dat wordt gebruikt in een woord, bijvoorbeeld 3 als E, of de klank van een cijfer in het Nederlands dat wordt gebruikt, bijvoorbeeld ‘w88’ voor ‘wachten’.

Ook acroniemen voor woorden, smilies en afkortingen kun je op het internet in chatboxen, forums en sociale netwerksites tegen komen. Deze acroniemen voor woorden, afkortingen en smilies worden soms ook gezien als behorend bij breezertaal. Breezertaal kent geen alom geaccepteerd woordenboek en grammatica, hierdoor kan er geen complete definitie gegeven worden van het begrip ‘breezertaal’. Bij gebrek aan definitie is het niet duidelijk welke woorden wel of niet bij deze taal horen en wat de juiste schrijfwijze voor een woord is. Verschillende mensen die in breezertaal schrijven, gebruiken het zelfde woord met een andere schrijfwijze. Op sommige sociale netwerksites (o.a. girlsonly.nl, hunkz.nl en gurlz.nl) is het gebruik van breezertaal verboden. Dit betekent echter niet dat het op die sites niet toegepast wordt.

Het zal moeilijker zijn associatie analyse toe te passen op teksten die geschreven zijn in breezertaal en elke andere taal waarin niet iedereen dezelfde manier van schrijven voor één woord gebruikt. Omdat één item op meerdere manieren geschreven kan worden, kan dit invloed hebben op de associatieregels. Hierdoor kunnen de confidence en support lager liggen dan het in werkelijkheid is.

Wanneer verschillende talen in dezelfde dataset gebruikt worden, moet hier rekening mee gehouden worden. Hierbij kan aan officiële talen (nederlands, engels) en niet officiële talen (breezertaal, straattaal) gedacht worden. Bij meerdere talen in één dataset worden woorden niet automatisch goed gegroepeerd, omdat het zelfde woord in meerdere talen geschreven is. Daarnaast verschillen homoniemen en synoniemen per taal, wat groepering nog moeilijker maakt.

9.6.2 Homoniemen

In paragraaf 10.4.5 wordt beschreven hoe woorden gegroepeerd worden om samen voldoende woordfrequentie te behalen om associatie analyse toe te passen. Het samenvoegen van homoniemen is echter wat lastiger. Homoniemen zijn woorden die dezelfde spelling, maar een verschillende betekenis hebben. Bijvoorbeeld het woord ‘stel’, dat kan gebruiken worden als: ‘wij zijn een stel’, ‘stel dat dit gaat gebeuren’ of ‘een stel studenten’.

Homoniemen zijn lastig samen te voegen, wanneer de context van de woorden niet duidelijk is. Het feit dat de homoniemen afkomstig zijn van een sociale netwerksite is niet de reden voor het moeizame samenvoegen. Moeilijkheden bij het samenvoegen van homoniemen zijn afkomstig van het gebrek aan context. Het gebrek aan context van losse woorden wordt veroorzaakt door de werkwijze van dit experiment. Gevolg is dat homoniemen een uitdaging vormen bij het groeperen van woorden.

9.6.3 Keuzevelden en standaardwaarden

Bij het aanmaken van een profiel, dient men een aantal verplichte velden in te vullen. Een groot deel van de velden op een sociale netwerksite bestaat uit keuzelijsten. De volgorde waarin de keuzeopties opgesomd staan is van groot belang. Dit kan er namelijk toe leiden dat mensen denken dat dit de heersende norm op de site is, waardoor zij zich daaraan willen voldoen. Uit een artikel van Boyd [O2] blijkt dat het willen voldoen aan de heersende norm een reden kan zijn om te liegen. Er wordt dan gekozen voor hetgeen zij op basis van het formulier als de heersende norm zien. Ook geeft Boyd aan dat mensen die verplichte keuzevelden liever niet invullen (bijvoorbeeld omwille van privacy), meestal de eerst genoemde optie kiezen [O2].

Wanneer standaardwaarden zijn ingevuld in de velden, zoals bijvoorbeeld bij Ikku.nl het geval is, kan dat er toe leiden dat mensen de standaardwaarden zo laten staan, ook wanneer dat niet de waarheid is. Ik vermoed dat men dit vooral doet wanneer men geen zin heeft de tijd te nemen het formulier in te vullen.

Zoals hierboven beschreven is, zijn er verschillende redenen om voor de eerst genoemde optie of de standaardwaarde te kiezen, wanneer dit niet klopt met de werkelijkheid. Ik vermoed dat dit vooral gebeurd bij verplichte velden en nauwelijks aan de orde is bij niet verplichte velden. Men kan er immers voor kiezen de niet verplichte velden leeg te laten. De waarden ingevuld in de verplichte velden zijn hierdoor waarschijnlijk minder betrouwbaar.

9.3 Keuze algoritme

Het experiment dat uitgevoerd is past datamining toe op profielen van sociale netwerksites. Zoals beschreven, is er sprake van web content mining. Er zijn factoren die een beperking opleggen aan het te kiezen algoritme. In deze paragraaf staan de hiervan afgeleide voorwaarden beschreven en de keuze die aan de hand van deze voorwaarden gemaakt is.

9.3.1 Voorwaarden algoritme

Het algoritme werd gekozen op basis van de volgende eisen:

- Het algoritme dient associatieve of causale regels als output te bieden. De resulterende regels kunnen dan gebruikt worden om advertenties persoonsgericht te kunnen plaatsen.
- Het algoritme moet een unsupervised learning algoritme zijn.
- Het algoritme dient zichzelf al bewezen te hebben in de wetenschappelijke wereld.
- Het algoritme moet opgenomen zijn in een datamining applicatie, die gebruiksvriendelijk is en aansluit op de dataset. Er dient voor dit onderzoek beschikking kunnen worden verkregen over het programma.

Een voorwaarde die aan het algoritme wordt gesteld is dat het associatieve of causale regels als output biedt. Causale regels zijn regels die een oorzakelijk verband aangeven, in de vorm van: als X dan Y . Associatieve regels zijn regels als: Als X gebeurd is het waarschijnlijk dat ook Y gebeurd, zonder aan te geven of X, Y of een andere variabele de oorzaak is. In paragraaf 9.4.1 kunt u meer lezen over associatieregels. Het algoritme dient regels als output te bieden omdat regels (indien deze van voldoende kwaliteit zijn) goed gebruikt kunnen worden om advertenties persoonsgericht te plaatsen. Een groepering van onderzoeksobjecten is bijvoorbeeld minder bruikbaar.

In de dataset van sociale netwerksites zie ik geen logisch onderwerp om de variabelen daarvan als enige mogelijke consequent te gebruiken. Ook wordt de informatievraag van de doelgroep (in dit geval adverteerders) niet duidelijk in één onderwerp omvat. Hieruit blijkt dat essentiele elementen ontbreken om te kiezen voor supervised algoritmes binnen dit experiment. Een supervised algoritme geeft namelijk een resultaat waar variabelen op één onderwerp de consequent vormen. Er is geen sterk argument om een supervised learning algoritme te gebruiken en een consequent om daarvoor te gebruiken steekt er ook niet uit, daarom kan het beste een unsupervised learning algoritme gebruikt worden.

Het algoritme dient zichzelf al bewezen te hebben. De slagingskans van het experiment is groter als zeker is of dat algoritme goed werkt. Bij tegenvallende resultaten kan beter geconcludeerd worden dat datamining op profielen van sociale netwerksites op de in dit onderzoek gebruikte wijze niet nuttig is, omdat de kans dat de tegenvallende resultaten aan het algoritme liggen zo veel mogelijk uitgesloten zijn. Bij positieve resultaten geeft een bewezen algoritme meer zekerheid dat dezelfde methode bij een andere dataset ook goede resultaten bereikt.

Aanwezigheid van het algoritme in een datamining applicatie is essentieel, aangezien het om praktische redenen niet binnen dit onderzoek past een algoritme te implementeren tot een applicatie. De applicatie die het algoritme bevat, dient gebruikersvriendelijk te zijn en de dataset dient hierin geladen te kunnen worden zonder veel aanpassingen te moeten doen aan de vorm van de dataset. Een programma dat veel gebruikt wordt heeft de voorkeur aangezien het programma zichzelf dan bewezen heeft, er voldoende gebruikersdocumentatie zal bestaan en de kans op cruciale fouten in de programmatuur verkleind wordt.

9.3.2 Mogelijke algoritmes

Associatie analyse is de meest logische keus wanneer de wens is associatieve regels te zoeken in een dataset. Associatie analyse is een vorm van datamining die door middel van algoritmes associatieve regels oplevert. Associatie analyse algoritmes zijn altijd unsupervised. Wanneer men associatie regels als resultaat wil hebben, wordt regelmatig eerst gebruik gemaakt van clustering, alvorens men associatie analyse toepast.

Hoewel associatie analyse de meest voor de hand liggende keuze is gaan we ook kijken naar andere vormen van datamining die associatieve of causale regels op kunnen leveren.

Het probleem van het vinden van associatie regels heeft gelijkenissen met het zoeken naar classificatieregels, het ontdekken van causale regels (o.a. Bayesian Belief Networks) en het leren van logische definities [O50].

Het generen van associatieve regels kan door middel van classificatie. Men kiest dan vooraf de mogelijke variabelen op een onderwerp voor de consequent en gaat op zoek naar de daarbij passende antecedent. Classificatie is een supervised algoritme, aangezien vooraf de variabelen voor de consequent bepaald worden. Een van de voorwaarden aan het algoritme voor dit experiment is dat het algoritme unsupervised moet zijn. Classificatie voldoet niet aan die voorwaarde en is daardoor niet geschikt voor dit experiment.

Bayesian Belief Networks (BBN) leveren regels op. Hoewel dit algoritme door Tan et al. [O38] in een lijst classificatie algoritmes wordt weergegeven, kan het gebruikt worden om associatieve regels te generen. Deze regels zijn dan niet alleen associatief, maar geven soms ook een causaliteit aan. Een nadeel van BBN is dat het veel tijd kost een netwerk te construeren [O38]. De tijd die het kost, maakt het gebruik van Bayesian Belief Networks duur. Daarnaast zijn Bayesian Belief Networks een omslachtige keuze wanneer dit moet resulteren in regels aangezien eerst een netwerk gemaakt dient te worden.

9.3.3 Keuze algoritme

Uiteindelijk blijkt dus dat associatie analyse en de Bayesian Belief Networks overblijven als reële opties voor dit onderzoek.

Artikelen waarin associatie analyse wordt toegepast op profielen met persoonlijke eigenschappen van mensen heb ik dusdanig weinig kunnen vinden binnen de onderzoeksperiode, dat ik aanneem dat associatie analyse weinig op die manier wordt toegepast. Dit onderzoek levert daarom een bijdrage op de vraag of associatie analyse met goede resultaten kan worden toegepast op profielen van mensen. Geralateerde artikelen die wel voorhanden beschrijven onderzoek naar associatie analyse op persoonlijke informatie, namelijk een onderzoek naar geweldplegers [O44] en een onderzoek naar klanten van een bank [O45]. Bij beide onderzoeken is Apriori gebruikt als algoritme.

Ook over de toepassing van Bayesian Belief Networks op persoonlijke informatie heb ik weinig artikelen kunnen vinden binnen de onderzoeksperiode. Artikelen die ik wel heb kunnen vinden gingen o.a. over het ontwikkelen van deze netwerken voor eigenschappen van gebruikers van virituele communities [O51]. In dit artikel wordt Sociale netwerk analyse gebruikt om aan de hand daarvan een profiel van de gebruiker op te stellen. Van deze profielen wordt vervolgens een Bayesian Belief Network geabstraheerd.

Hoewel Bayesian Belief Networks en associatie analyse naar mijn mening beiden geschikte vormen van datamining zouden zijn voor dit experiment, heb ik gekozen voor associatie analyse omdat Bayesian Belief Networks duurder en omslachtiger zijn.

9.4 Associatie analyse

In de voorgaande paragraaf is de keuze gemaakt voor associatie analyse. Nu zal worden toegelicht hoe associatie analyse in zijn werk gaat, hoe de algoritmes werken en zal de keuze voor een specifiek algoritme binnen associatie analyse toegelicht worden.

9.4.1 Associatieregels

Het genereren van associatieve regels kan op verschillende manieren. Een associatie regel is een expressie $X \rightarrow Y$, waar X en Y sets van items zijn. De itemset voor de pijl (in dit geval X) wordt antecedent genoemd en de itemset achter de pijl (in dit geval Y) wordt consequent genoemd. Gegeven een database D gevuld met transacties, waar elke transactie $T \in D$ een set items is. $X \rightarrow Y$ betekent dat wanneer de transactie T X bevat, T ook Y bevat. De kwaliteitsmaten geven vervolgens aan hoe vaak deze regel op gaat [O52]. De items die in een regel aan de orde komen worden gezamenlijk de itemset genoemd. Binnen dit experiment wordt de informatie behorende bij een identiteit gezien als transactie. Eigenschappen als leeftijd, woonplaats, hobby's en dergelijke worden gezien als items.

9.4.2 Kwaliteitsmetingen

Support en confidence zijn veelgebruikte kwaliteitsmetingen om de kwaliteit van een associatieregel te schatten. De reden hiervoor is dat zij eigenschappen hebben waarmee het aantal te gebruiken itemsets en het aantal potentiële regels efficiënt gereduceerd kunnen worden. Dit maakt dat support en confidence erg geschikt zijn voor gebruik in associatie analyse tools, waar zij dan ook veel toegepast worden. De aanwezigheid van support en confidence in veel tools is de aanleiding deze metingen hieronder kort toe te lichten

De support meet de ondersteuning van de regel door de data. De support van: X Y is: (σ(X U Y)) / N, waar N het totaal aantal transactieregels is. Dus als het aantal dat X en Y bevat 2 is en het totale aantal transacties 5 is, dan is de support 2/5=0,4. De support is belangrijk omdat deze meetwaarde regels eruit filtert die waarschijnlijk veroorzaakt zijn door toeval. Daarnaast is een regel met een lage support over het algemeen niet interessant vanuit bedrijfs- of marketing perspectief [O38], omdat de meeste bedrijven zich richten op massabereik.

De confidence geeft de kans aan dat Y voorkomt als X voorkomt. Dit is de betrouwbaarheid van de regel. De confidence van: X → Y is: (σ(X U Y)) / σ X. Dus als het aantal transacties waarvoor X→ Y geldt 2 is en het aantal transacties waar Y in voorkomt 3 is, dan is de confidence 2/3=0,67. De confidence geeft aan of X → Y de beste regel is die uit de itemset {X, Y} gehaald kan worden [O38]. Het zou ook kunnen dat de regel Y → X beter is dan de regel X → Y.

Support en confidence hebben ook hun minpunten, welke hieronder uiteengezet worden. Vanwege hun minpunten, bestaan er ook andere kwaliteitsmetingen, bijvoorbeeld de Correlatie, Kappa, Interest, Cosinus en Jaccard [O38].

Regels waarvan de support laag is, kunnen interessant zijn als zij een hoge confidence hebben en het doel van de associatie analyse is om regels te genereren aan de hand waarvan mensen individueel aangesproken worden. Lui et al. [O53] hebben succesvol een methode gebruikt met multiple minimum support om dit probleem te verkleinen. In deze methode is de support van een itemset gedefinieerd als de minimumsupport van alle items in de itemset.

Confidence kan als kwaliteitsmeting een vertekend beeld geven. Wanneer zou blijken dat de confidence van $X \rightarrow Y$ hoog is, bijvoorbeeld 0,75 (dit betekent dat de kans dat Y voorkomt 75 procent is, wanneer X voorkomt), maar de kans dat Y voorkomt onafhankelijk van andere items hoger is, bijvoorbeeld 80 procent, maakt X de kans kleiner dat Y voorkomt. Dan is de aanwezigheid van X dus geen goede aanwijzing voor de aanwezigheid van Y en heeft de regel dus geen nut wanneer men het voorkomen van Y wil vergroten [O38].

9.4.3 Definitie associatie analyse

Datamining met als doel het vinden van associatieve regels werd in 1993 geïntroduceerd door Agrawal et al. [O54]. Associatie analyse wordt ook wel afhankelijkheidsanalyse [O55] of boodschappenmandjes-analyse genoemd [O57].

9.4.4 Werking van associatie analyse algoritmes

Associatie algoritmes gebruik makend van confidence en support, werken over het algemeen als volgt: Het algoritme krijgt de opdracht “Gegeven een set transacties T, zoek alle regels waar voor geld dat de support groter of gelijk is aan de vooraf bepaalde minimum support en de confidence groter of gelijk is aan de vooraf bepaalde minimum confidence”. Een brute kracht aanpak om associatie regels te zoeken is door support en confidence te berekenen voor elke mogelijke regel. Deze regel kost erg veel tijd en inspanning omdat er zelfs met een betrekkelijk klein aantal verschillende items in de dataset er al erg veel mogelijke regels ontstaan. Het totale aantal mogelijke regels uit een dataset met d items is: $R = 3^d - 2^{d+1} + 1$ [O38].

De standaard strategie in veel algoritmes voor het genereren van associatieve regels is verwerkt is de volgende [O38]:

- Genereren van frequente itemsets. Alle itemsets die voldoen aan de minimale supportwaarde, worden frequente itemsets genoemd.
- Genereren van regels, het doel hiervan is om van alle regels de confidence te extraheren uit de frequente itemsets die bij de vorige stap zijn gevonden. Deze regels worden sterke regels genoemd.

Het reduceren van de complexiteit van het genereren van frequente itemsets kan door het aantal kandidaat itemsets te reduceren (dwz een aantal kandidaat itemsets uitsluiten zonder hun support waardes te meten), of door het aantal vergelijkingen van kandidaat itemsets met transacties reduceren (bijvoorbeeld door middel van een hash tree) [O38]. Het zoeken van frequente itemsets kan door middel van verschillende strategieën, er is niet één optimale strategie. De beste strategie voor het genereren van frequente itemsets is afhankelijk van de dataset en de strategie voor het genereren van regels.

Omdat de support voor $X \rightarrow Y$ gelijk is aan de support voor $Y \rightarrow X$ kan dit gegeven gebruikt worden om het aantal te bekijken regels te beperken. De supportmeting is kijkt naar de frequentie van de combinatie van items, dit is ook de reden dat support hetzelfde is wanneer je dezelfde itemsets in je regel hebt staan. Om de support te weten van beide regels, hoeft deze maar voor één van de twee uitgerekend te worden. Het uitrekenen voor één regel en gebruiken van de uitkomst voor beide regels maakt het algoritme efficiënter. Regels gebaseerd zijn op infrequente itemsets kunnen verwijderd worden uit de lijst potentiële associatieregels omdat zij niet zullen voldoen aan de minimale support [O38].

Uit elke frequente k-itemset (waarbij k staat voor het aantal items) kunnen $2^k - 2$ associatieregels gegenereerd worden, wanneer regels met lege consequenten of antecedenten uitgesloten worden. Een associatieregel kan geëxtraheerd worden van een itemset door de itemset ‘Z’ op te delen in twee niet lege itemsets, X en Z - X, zodat $X \rightarrow Z - X$ voldoet aan de minimum confidence. Alle regels die op die manier gegenereerd worden, voldoen al aan het support minimum omdat ze gegenereerd zijn uit een frequente itemset [O38].

Algoritmes kunnen gekarakteriseerd worden aan de hand van de strategie waarmee de data (groep transacties) onderzocht wordt en de strategie om de supportwaarden van de itemsets te bepalen.

9.4.5 Associatie analyse algoritmes

Een aantal associatie analyse algoritmes zijn:

- Apriori [O56]
- Apriori TID [O50]
- AprioriHybryd [O50]
- CBA-RG [O58]
- DIC [O59]
- E-Apriori [O60]
- Eclat [O62]
- EH-Apriori [O61]
- FDM [O63]
- FP-Tree [O64]
- Matrix [O65]
- Predictive Apriori [O66]
- Prices [O67]
- SETM [O61]
- TAR2 [O68]
- Tertius [O69]
- Titanic [O70]

Bovenstaande lijst is verre van compleet, alleen een aantal associatie analyse algoritmes die ik tegen ben gekomen in de literatuur zijn hierin opgenomen. Apriori is het meest gebruikte algoritme om d.m.v. datamining associatieve regels te verkrijgen [O71]. Veel van bovenstaande algoritmes zijn gebaseerd op dit algoritme. Dit geld overigens niet alleen voor de algoritmes waar Apriori in de naam verwerkt is. In veel artikelen waarin een nieuw algoritme geïntroduceerd wordt, worden de resultaten van dit algoritme vergeleken met het Apriori algoritme. Door het Apriori algoritme te gebruiken binnen een onderzoek, heeft men de keuze gemaakt voor een bewezen algoritme en de resulterende meetwaarden van het experiment kunnen eenvoudig vergeleken worden met andere onderzoeken. Andere hierboven genoemde algoritmes worden op een enkeling na (bijvoorbeeld het eerste associatie analyse algoritme: AIS) nauwelijks genoemd in meerdere artikelen. Ook wordt afgaande op de voor dit onderzoek doorgenomen literatuur over nieuwe algoritmes, geen van de hierboven genoemde algoritmes zo vaak vergeleken met een nieuw algoritme als Apriori.

De voorwaarden die in paragraaf 9.3.1 gesteld werden aan een algoritme, zijn de volgende:

- Het algoritme dient associatieve of causale regels als output te bieden. Deze regels kunnen dan gebruikt worden om advertenties persoonsgericht te kunnen plaatsen.
- Het algoritme moet een unsupervised learning algoritme zijn.
- Het algoritme dient zichzelf al bewezen te hebben in de wetenschappelijke wereld.
- Het algoritme moet aanwezig zijn in een datamining applicatie, die gebruiksvriendelijk is en aansluit op de dataset. Er dient voor dit onderzoek besschikking kunnen worden verkregen over het programma. Een programma dat veel gebruikt wordt geniet de voorkeur.

Associatie analyse algoritmes bieden associatieve regels als output en zijn unsupervised. Alle in de opsomming genoemde algoritmes voldoen dus aan de eerste twee voorwaarden. Zoals we in het begin van deze paragraaf gezien hebben, is Apriori een algoritme dat zichzelf al bewezen heeft, terwijl dit voor de meeste andere algoritmes niet zo is. Volgens Becquet et al. [O72] bevatten veel datamining tools een implementatie van het Apriori algoritme, waaronder commerciële pakketen als Clementine (huidige naam: SPSS PASW Modeler) en gratis academische software pakketten als Weka [O73] en CBA [O74].

9.5 CBA-RG

Liu et al. [O74] hebben het appriori algiritme [O56] in het programma CBA geïmplementeerd en CBA-RG genoemd. De afgelopen jaren is CBA nog gebruikt in wetenschappelijk onderzoek [O75] [O76] [O77] [O78]. De afkorting staat voor Classified Based Associations. CBA bevat naast de regel generator (CBA-RG) een implementatie van een classificatiealgoritme (CBA-CB genoemd) [O74].

CBA-RG zoekt allereerst alle itemsets die een support hebben boven de gestelde minimum support. Zoals eerder genoemd is de support het aantal transacties waarin de itemset aanwezig is ten opzichte van de totale dataset, uitgedrukt in een percentage.

Het algoritme genereert frequente itemsets door meerdere passes door de dataset te maken. De eerste keer berekend CBA-RG de support van individuele items en bepaald of het item voldoet aan de minimum support. In elke volgende pass, worden een aantel items gepakt die frequent zijn bevonden in vorige pass, die de seed set genoemd worden. CBA-RG gebruikt de seedset van de vorige pass, om nieuwe kandidaat itemsets te genereren. De daadwerkelijke support van deze itemsets worden bepaald wanneer het algoritme door de dataset heen loopt. Wanneer het algoritme door de gehele dataset is gelopen, bepaald het de frequente kandidaat itemsets. Vanuit de frequente itemsets worden kandidaat regels gegenereerd. Van alle mogelijke regels die gegenereerd kunnen worden uit de frequente itemsets, wordt de regel met de hoogste confidence gekozen waneer regels dezelfde set antecedenten hebben. Wanneer meerdere items dezelfde hoogste confidence hebben, wordt er willekeurig een regel gekozen [O79].

F_1 = {large 1-ruleitems};
Car_1 = genRules(F_1)
for (k =2; $F_{k-1} \neq \emptyset$; k++) do
    C_k = candidateGen(F_{k-1});
    For each data case d Є D do
        C_d = ruleSubset (C_k, d);
        For each candidate c Є C_d do
            c.condsupCount++;
            if d.class = c.class then c.rulesupCount++;
        end
    end
    Fk = {c Є C_k | c.rulesupCount >= minsup};
    CAR_k = genRules(F_k);
End
CARs = U_k CAR_k;

Werking CBA-RG / Apriori [O79]

Hierboven is de werking van het algoritme beschreven. In deze beschrijving staat k voor het aantal items, D voor de dataset, d voor een itemset, C staat voor de kandidaat itemset en CAR voor de gegenereerde regels.

Binnen literatuur over CBA-RG worden de set mogelijke regels CARS genoemd (class association rules). De consequent wordt in literatuur over CBA soms klasse genoemd, als verwijzing naar classificering. Associatieregels werken niet met klasses. Binnen de literatuur over dit algoritme worden de consequenten zo genoemd omdat CBA-RG regelmatig gebruikt wordt in combinatie met CBA-CB. CBA-RG levert regels op met allerlei consequenten, er wordt niet gevraagd een klasse aan te wijzen of iets dergelijks. Omdat binnen dit experiment CBA-CB niet gebruikt wordt, worden benamingen als CARS en klasse niet gebruikt in de verdere beschrijving van dit experiment. CBA-RG is prima te gebruiken zonder aanwezigheid van klasses en zonder het te combineren met CBA-CB.

9.5.1 Apriori Algoritme

Apriori is het eerste associatie analyse algoritme dat gebruik maakte van het uitsluiten van itemsets gebaseerd op support, om daarmee systematisch controle te krijgen over de exponentiele groei van kandidaat itemsets [O38].

Brute kracht:

$$\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = 6 + 15 + 20 = 41$$

Apriori:

$$\binom{6}{1} + \binom{4}{2} + 1 = 6 + 6 + 1 = 13$$

Hierboven staat de werking van het Apriori algoritme beschreven. De effectiviteit van de strategie van het Apriori algoritme kan bewezen worden door te kijken naar het aantal kandidaat itemsets dat gegenereerd is. Een brute kracht strategie van het genereren van itemsets (tot een grote van 3) geeft 41 kandidaat itemsets, terwijl dit aantaal daalt tot 13 kandidaten met het Apriori Algoritme [O38].

Zoals ook al in paragraaf 9.4.5 genoemd werd, zijn er veel algoritmes die gebaseerd zijn op het Apriori algoritme. De wijziging is voor sommige datasets een verbetering, maar in de meeste gevallen niet voor alle soorten datasets. Deze algoritmes zijn nog niet voldoende gevalideerd op voldoende verschillende datasets om ze als bewezen algoritmes te kenmerken.

10 Aanpak experiment Associatie analyse

In dit hoofdstuk wordt de aanpak van het associatie analyse experiment beschreven. Binnen dit experiment wordt een algoritme toegepast op data uit profielen van een aantal sociale netwerksites. Vervolgens worden dmv kwaliteitsmaten resulterende associatieregels verwijderd en worden de voor marketing nuttige associatieregels geselecteerd.

10.1 Onderzochte sociale netwerksites

Er is contact opgenomen met veertig sociale netwerksites met de vraag of zij medewerking wilden verlenen. Voor het datamining-experiment hebben twee sites hun medewerking toegezegd: Ikku.nl en Girlsonly.nl. Gezamenlijk hebben zij data van 70.355 profielen aangeleverd.

Lui en Maes [O27] geven aan dat het analyseren van gegevens van meerdere sociale netwerksites als voordeel heeft dat de gebruikte site minder invloed heeft op de uitkomst van het onderzoek. Als nadeel geven zij aan dat er een overlap is wat betreft gebruikers. Dit heeft als gevolg dat die gebruikers meer invloed hebben op de uiteindelijke resultaten. Wat dit betekend voor dit experiment, staat beschreven in paragraaf 6.7.1.

In dit experiment zijn geen gegevens van gebruikers meegenomen die uniek zijn voor een gebruiker (bijvoorbeeld naam of e-mailadres). Deze gegevens zijn nodig om gebruikers op te kunnen sporen die dubbel in de dataset voorkomen. Om de privacy van gebruikers te beschermen, is de beheerders gevraagd een geanonimiseerde dataset aan te leveren. Hierdoor konden dubbele leden niet opgespoord worden en kon er dus geen aanpassing plaatsvinden om ervoor te zorgen dat een gebruiker slechts één keer voorkomt. Hierbij wordt de kanttekening geplaatst, dat leden wanneer zij op één site meerdere profielen hebben, zij zich meestal niet met hetzelfde e-mailadres kunnen inschrijven en mogelijk ook onder een andere naam/nickname ingeschreven staan. Uit paragraaf 8.3.1 blijkt namelijk dat anonimiteit een veel voorkomende reden is voor het bezit van meerdere profielen.

10.2 Input data

Zoals te zien is in hoofdstuk 7, is de input van associatie analyse in de vorm van transacties. In dit experiment zal de data uit profielen als transactie worden gezien. Waardoor informatie over één identiteit wordt in dit experiment gezien als één transactie. Hierbij dient wel de kanttekening geplaatst worden dat, zoals we gezien hebben in paragraaf 6.7.1, één persoon meerdere identiteiten kan hebben op een sociale netwerksite, waardoor de door hem opgegeven informatie opgeslagen is binnen twee 'transacties'.

Omwille van de privacy van de gebruikers van de meewerkende sites, wordt niet alle data op een profiel gebruikt voor dit experiment. Privacygevoelige informatie als voornaam, achternaam, nickname, woonplaats en emailadres worden niet gebruikt binnen dit experiment.

De volgende informatie wordt gebruikt dit experiment:

- Site {Ikku.nl, Girlsonly.nl}
- Geboortejaar
- Provincie
  {Gelderland, Limburg, Friesland, Utrecht, Zuid-Holland, Noord-Holland, Overig, Overijsel, Noord-Braband, Groningen, Drente, Zeeland, Flevoland}
- Opleiding {vwo, havo, mavo, vmbo-t, vmbo-g, vmbo-k, vmbo-b, vbo, anders}
- Geslacht {man, vrouw}
- Geaardheid {hetero, homo, biseksueel}
- Relatie {ja, nee}
- Huidskleur {blank, bruin, zwart}
- Kleur ogen {bruin, blauw}
- Losse tekst (meer hierover in paragraaf 10.4)
- Sterrenbeeld

Omdat CBA-RG geen continue variabelen aan kan, zijn de continue waarden omgezet naar binaire variabelen. Zo is het geboortejaar gebruikt als binaire variabele door aan te geven of dit voorkomt in de dataset.

10.3 Aanpak van moeilijkheden

10.3.1 Taalgebruik

Zoals in paragraaf 9.6 is besproken, wordt op veel sociale netwerksites niet alles volgens de Nederlandse spelling geschreven. Bij een korte inventarisatie van de door de sociale netwerksites aangeleverde data, werd al duidelijk dat niet alles wordt geschreven volgens de Nederlandse spelling.

De volgende zinnen zijn afkomstig van profielen en geven een indruk van zinnen gebruikt op sociale netwerksites. Deze zinnen bestaan uit een mengeling van breezertaal, Engelse woorden en foutief en correct gespelde Nederlandse woorden:

Citaat	**Betekenis**
“welcome op muh pro…”	Welkom op mijn profiel
“End d0nt c0py my pr0!!!”	En kopieer mijn profiel niet!
“Wil jh helpee?	Wil je me helpen?
“Mail me dn dr te zge hoe foto’s toevoege!”	Mail me dan om te zeggen hoe ik foto’s toe kan voegen.

Het is mogelijk dat ook Turkse, Marokkaanse of Surinaamse woorden op Nederlandse sociale netwerksites voorkomen, aangezien woorden uit die talen ook in straattaal voorkomen [O80].

Het gebruik van verschillende spellingswijzen zal vooral voorkomen in de losse tekst die gebruikers op hun profiel kunnen zetten. Het gebruik van verschillende spellingwijzen in de losse tekst wordt aangepakt door groeperingen van woorden handmatig uit te voeren. Meer over de groepering van woorden staat beschreven in paragraaf 10.4.5.

De overige data wordt door gebruikers in velden ingevuld. Veel van de velden waarvan de data in dit experiment wordt gebruikt zijn keuzevelden. Het gebruik van keuzevelden voorkomt verschillende spellingswijzen voor het zelfde woord.

10.3.2 Homoniemen

In paragraaf 10.4.5 wordt beschreven hoe woorden gegroepeerd worden om samen voldoende woordfrequentie te behalen om associatie analyse toe te passen. Het correct samenvoegen van homoniemen met andere woorden is echter wat lastiger. Homoniemen zijn woorden die dezelfde spelling, maar een verschillende betekenis hebben.

Zoals in hoofdstuk 8 maakt gebrek aan context het moeilijk homoniemen samen te voegen. De oorzaak van gebrek aan context is de werkwijze in dit experiment, de sociale netwerksites zijn geen veroorzakende partij. Er zijn manieren om de context te kunnen gebruiken waardoor homoniemen waarschijnlijk ook beter te categoriseren zijn. Deze zijn niet gebruikt in dit experiment, maar wel besproken in hoofdstuk 14. Binnen dit experiment is bij het groeperen van woorden er voor gekozen homoniemen alleen samen te voegen met andere woorden, wanneer aannemelijk gemaakt kan worden dat er één betekenis is die door de jongeren op de site waarschijnlijk als enige betekenis gebruikt wordt. In alle overige gevallen is het woord niet samengevoegd met andere woorden. Dat kan er voor hebben gezorgd dat woorden de minimum frequentie niet gehaald hebben, welke gesteld is in paragraaf 10.4.4.

10.3.3 Keuzevelden en standaardwaarden

In paragraaf 6.3.4 wordt beschreven dat verplichte keuzevelden en invulvelden waar al een standaardwaarde is ingevuld, waarschijnlijk minder betrouwbaar zijn. Omdat in hoofdstuk 7 en 8 al gekeken is naar de betrouwbaarheid van data op profielen, zal dat in dit experiment niet apart worden meegenomen. Na uitvoering van het experiment zijn de resultaten beschreven in hoofdstuk 11, geanalyseerd met de resultaten uit hoofdstuk 8. Het databetrouwbaarheidsonderzoek resulteert op die manier in extra duidelijkheid over de zekerheid van de associatieregels. De combinatie van die twee experimenten is beschreven in hoofdstuk 12.

10.4 Losse tekst

Een gebruiker voert voor zijn profiel een aantal velden in en heeft daarnaast de besschikking over een vrij in te vullen ruimte. Vanuit de losse tekst is een lijst samengesteld van veelgebruikte woordgroepen. De wijze waarop dit gedaan is wordt op de volgende pagina's uiteengezet. Na het samenstellen van de woordgroepenlijst, is per profiel en per woordgroep gekeken of een woord uit de woordgroep voorkomt in het profiel. Dit al dan niet voorkomen van een woordgroep in het profiel, vormde een binaire variabele.

Allereerst zijn een aantal handelingen uitgevoerd om de data te zuiveren. Vervolgens zijn een aantal selectie criteria gebruikt om de woorden te selecteren. Daarna zijn een aantal handelingen uitgevoerd waarin het aantal voorkomens van woorden die op elkaar lijken of het zelfde onderwerp hebben, bij elkaar opgeteld worden. Ten slotte is een gedeelte van de geselecteerde woorden verwijderd, omdat zij niet vaak genoeg voorkwamen of weinig betekenis hadden.

Voor het selecteren van de te gebruiken woordgroepen zijn de volgende specifieke stappen gevolgd (in de hier weergegeven volgorde):

Datazuivering:

1. HTML verwijderen d.m.v. HTML filter
2. Aantal voorkomens van woorden tellen
3. Losse streepjes en underscores buitensluiten
4. Losstaande getallen buitensluiten (incl. jaartallen en rangtelwoorden)

Selectie van woorden

5. Woorden met een woordfrequentie lager dan 300 buitensluiten
6. Zelfstandige naamwoorden selecteren
7. Werkwoorden selecteren
8. Bijvoeglijke naamwoorden selecteren

Samenvoeging van woordfrequenties

9. Woordfrequentie van gelijke woorden samenvoegen.
10. Woordfrequentie van meervoud en enkelvoud samenvoegen
11. Woordfrequentie van synoniemen samenvoegen.
12. Woordfrequentie van woorden met een sterk verband samenvoegen
13. Woordfrequentie van woorden die een eigenschap zijn van een ander woord samenvoegen.

Verwijdering van woorden

14. Woordengroepen verwijderen met een woordfrequentie lager dan 704 (gebaseerd op supportgrens)
15. Te algemene en nietszeggende woorden verwijderen.

Woorden groeperen en verwijderen aan de hand van supportwaarde

16. Berekenen support van woorden
17. Woorden onder de supportgrens groeperen indien mogelijk
18. Woorden onder de supportgrens verwijderen
19. Parent-groepen aanmaken

Handeling één en twee zijn geautomatiseerd uitgevoerd. Handeling drie tot en met vijftien en zeventien tot en met negentien zijn handmatig uitgevoerd. Handeling zestien is geautomatiseerd uitgevoerd.

10.4.1 Datazuivering

De HTML is verwijderd omdat deze codes in elk profiel zit en over het algemeen weinig zegt over de persoon. De meeste stukken code komen voor in elk profiel. Waar HTML code profielspecifiek is, bevatten deze meestal informatie over de opmaak van het profiel (bijvoorbeeld lettertype of locatie van een afbeelding.

Op Ikku.nl heeft een gebruiker meer vrijheid wat betreft opmaak van het profiel dan op Girlsonly.nl. Op Ikku kan men door middel van HTML afbeeldingen en tekst op zijn profiel aanpassen. Op Girlsonly kan alleen de achtergrond kleur die men zelf ziet veranderd worden. Anderen zien dan echter de door henzelf ingestelde kleur. Waar kleuren gewijzigd kunnen worden en afbeeldingen kunnen worden toegevoegd kan dit iets zeggen over de persoon. De kleur kan de lievelingskleur van de persoon zijn en de bestandsnaam van de afbeelding kan een aanwijzing zijn voor wat er weergegeven wordt op de afbeelding. Dit hoeft echter niet het geval te zijn. De kleur kan ook om andere redenen gekozen zijn en de bestandsnaam van de afbeelding zal in veel gevallen ook niets zeggen over de afbeelding. Dit wil niet zeggen dat deze gegevens waardeloos zijn. Kennis over de kleur van de achtergrond, kan bijvoorbeeld meegenomen worden in de keuze van de kleuren in een advertentie.

Om praktische redenen, zal alle HTML uit de data gefilterd moeten worden, of een groot aantal gegevens uit de HTML gebruikt moeten worden. Aan de hand van onderstaande redenen is de keuze gemaakt zoveel mogelijk HTML code te verwijderen uit de losse tekst door middel van een HTML filter:

- Niet op alle sites is het toevoegen van eigen HTML of het wijzigen van opmaak van het profiel mogelijk.
- De meeste HTML is niet profielspecifiek.
- Het grootste deel van de profielspecifieke HTML tags zegt niet veel over eigenschappen van de persoon.
- Waar de HTML tags iets zouden kunnen zeggen over eigenschappen van de persoon, is vaak niet duidelijk wat de bijbehorende eigenschap zou zijn omdat de reden voor de keuze van bijvoorbeeld de achtergrondkleur niet duidelijk is. Dit maakt de waarde van deze tags onduidelijk. Zoals hierboven beschreven is het kennen van de achtergrondkleur niet waardeloos, maar om praktische redenen is besloten niet alleen één eigenschap te gebruiken die beschreven is in de HTML.
- Het is complex een onderscheid te maken tussen verschillende tags die bijvoorbeeld allemaal een kleur als eigenschap hebben, omdat de context niet altijd duidelijk is.

Vervolgens is het programma TextSTAT[2] gebruikt dat beschikbaar wordt gesteld door de vrije universiteit van Berlijn, om het aantal voorkomens van een woord te tellen. De frequentie van een woord wordt gebruikt om veelvoorkomende woorden op te selecteren.

[2] TextSTAT - Simple Text Analysis Tool, versie 4.42.0.0,
Free University of Berlin - Dutch Linguistics, 15 Januari 2008
http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

10.4.2 Selectie van woorden

Op basis van de woordfrequentie, is ervoor gekozen alleen de woorden te behouden die minstens 300 keer voorkomen. Aangezien het totale aantal woorden dat voorkomt in de profielen erg groot is, moest er een grens gesteld worden. De grens is op een frequentie van 300 gelegd, omdat het de woorden met een lagere frequentie een zeer kleine kans hebben na samenvoeging (zie 10.4.3) een frequentie behalen die hoger is dan 704. Dan zouden namelijk minstens 3 soortgelijke woorden voor moeten komen met een frequentie lager dan 300. Wanneer een soortgelijk woord een woordfrequentie hoger dan 704 heeft, worden de woorden in een later stadium samengevoegd, en gaat het woord met een voorkomen onder de 300 dus nog niet verloren.

De ene soort woorden geeft meer aan over de persoon als de ander. Lidwoorden, voornaamwoorden, voorzetsels en voegwoorden hebben bijvoorbeeld nauwelijks meer een betekenis wanneer zij uit de context gehaald worden. Er is voor gekozen alleen zelfstandige naamwoorden (zowel eigennamen als soortnamen), bijvoeglijke naamwoorden (zowel attributief als predicatief) en werkwoorden te selecteren.

De zelfstandige naamwoorden bevatten o.a. merk- en productnamen, namen van beroemdheden, gebruiksvoorwerpen, consumptieartikelen en woorden waaruit interesses en hobby's afgeleid kunnen worden. De bijvoeglijke naamwoorden bevatten o.a. kleuren, emoties en eigenschappen van mensen. De werkwoorden bevatten o.a. bezigheden en hobby's. Deze woorden geven hierdoor informatie over interesses, productgebruik, bezigheden en andere eigenschappen van jongeren.

10.4.3 Samenvoeging van woordfrequenties

Veel woorden in de selectie zullen op elkaar te lijken, veel met elkaar te maken te hebben of soms zelfs dezelfde betekenis te hebben. De woordfrequenties van deze woorden worden bij elkaar opgeteld, omdat het bij gebruik van twee bijna gelijke woorden, over hetzelfde onderwerp gaat. De woorden worden handmatig samengevoegd. Dit wordt handmatig gedaan omdat een geautomatiseerde oplossing verre van foutloos zal werken, omdat (zoals al beschreven in paragraaf 9.6) lang niet altijd correct Nederlands gebruikt wordt in de profielen.

Allereerst zijn woorden samengevoegd die precies hetzelfde betekenen, maar door ander hoofdletter gebruik, gebruik van een andere taal (breezertaal of engels) of verkeerde spelling niet gezien zijn als hetzelfde woord door het programma dat gebruikt is om het aantal voorkomens te tellen.

Vervolgens zijn de frequenties van het meervoud en enkelvoud van een woord bij elkaar opgeteld. Dit geeft een klein betekenisverlies aangezien iemand die het woord 'pretparken' gebruikt waarschijnlijk meer met pretparken in aanraking komt dan iemand die dit woord in het enkelvoud gebruikt. Veel gegevensverlies levert dit echter niet op; in de meeste gevallen geeft een verschil tussen enkel en meervoud niet direct een heel andere indruk van een persoon.

Synoniemen die voorkomen in de woordenlijst worden samengevoegd. Een voorbeeld hierbij zijn de woorden 'sterven' en 'overlijden'. Wanneer een van de woorden een homoniem is (dwz dat dit woord meerdere betekenissen heeft), is dit woord niet samengevoegd met een ander woord, behalve wanneer aannemelijk is dat de gebruikersgroep van de sociale netwerksite (Nederlandse jongeren) slechts één van deze betekenissen gebruikt en die betekenis overeen komt met die van het woord waarmee het woord samengevoegd kan worden.

Woorden met een woordfrequentie onder de 704 worden verder geanalyseerd. De grens betreffende de woordfrequentie is op 704 voorkomens gelegd omdat woorden met minder dan 704 voorkomens in een latere stap verwijderd zullen worden. Woorden met een frequentie onder de 704 worden samengevoegd wanneer zij die een eigenschap zijn van een ander woord of de betekenis veel lijkt op een ander woord, maar er niet gelijk aan is. De woorden 'auto' en 'automerk' zijn een voorbeeld van woorden die samengevoegd zijn omdat het ene woord een eigenschap is van het andere. Een voorbeeld van woorden die omwille van grote overeenkomstigheden in betekenis zijn samengevoegd, zijn 'zanger' en 'zangeres'.

In enkele gevallen zijn woordfrequenties niet bij elkaar opgeteld, omdat aannemelijk was dat de woorden altijd gezamenlijk voorkomen. Dit was bijvoorbeeld het geval bij de woorden 'Brad' en 'Pitt'. Deze woorden komen vaak samen voor (dankzij de acteur Brad Pitt) en hadden een bijna gelijke woordfrequentie. In dergelijke gevallen vormen de woorden wel een woordgroep, maar wordt het aantal voorkomens niet bij elkaar opgeteld.

10.4.4 Verwijderen van woordgroepen

In de vorige stappen zijn woordgroepen ontstaan, door het samenvoegen van woorden en woordfrequenties wanneer woorden die veel op elkaar leken of met elkaar te maken bleken te hebben. Een enkele keer bestaat een woordgroep uit maar één woord, wanneer dit woord niet op basis van de criteria genoemd in paragraaf 10.4.3 met andere woorden samengevoegd kon worden.

In deze stap worden woorden handmatig verwijderd op basis van twee criteria.
Wanneer de woordgroepen een gezamenlijke woordfrequentie hebben die lager is dan 704, worden deze woordgroepen verwijderd uit de selectie. De datasets die aangeleverd waren door de twee sociale netwerksites, omvatten gezamenlijk 70.356 profielen. Wanneer we uitgaan van een support van 0,01 betekend dat een woord in 704 profielen voor dient te komen om een support van 0,01 te behalen (0,01 x 70356 = 703,56). Een support van 0,01 wil zeggen dat een woord in minstens één procent van de profielen voorkomt. Er is echter geteld hoe vaak een woord voorkomt in de totale set losse data uit de profielen, zonder daarbij te kijken hoe vaak een woord per profiel voorkomt. Wanneer een persoon één woord heel vaak gebruikt, is bij een gelijke woordfrequentie, de support (het percentage profielen waarin het woord voorkomt) lager. Bij de woordfrequentiegrens van 704, betekend dit waarschijnlijk een klein deel van de woorden boven deze frequentiegrens, een support hebben die lager is dan 0,01.

Vervolgens wordt de woordenselectie geanalyseerd. Woorden die niets zeggen over de, eigenschappen, interesses of bezigheden van een persoon worden verwijderd. Woorden waarvan niet zeker is of deze informatief zijn of niet, worden behouden. Voorbeelden van verwijderde woorden zijn: aardig, beginnen, nieuw, nummer, pad, persoon, profiel, rechts, seizoen, simpel, soort, woord en zitten.

10.4.5 Woorden groeperen en verwijderen aan de hand van supportwaarde

De supportwaarde is berekend voor de geselecteerde woorden en woordgroepen. De woorden die wel het minimale aantal voorkomens hadden maar niet voldeden aan de support werden indien mogelijk gegroepeerd met andere woorden. De woorden waarmee zij gegroepeerd werden konden zowel onder of boven de supportgrens liggen. Woorden worden gegroepeerd aan de hand van het verband tussen onderwerpen.

Woorden met voldoende voorkomens konden onder de supportgrens komen te liggen, wanneer de personen die het woord in hun profiel hebben staan, dit woord meerdere malen gebruikten. Het aantal voorkomens werd geteld over de losse tekst van alle profielen samen. Bij de berekening van de support wordt gekeken of een woord al dan niet voorkomt in een profiel, het is hierbij niet van belang hoe vaak het woord voorkomt per profiel.

De grens die gesteld wordt bij het gebruik van het algoritme is een support van één procent. Woorden met een lagere support zullen dus niet voorkomen in associatieregels. Dit wil niet zeggen dat woorden met een support van minimaal één procent sowieso voor zullen komen in associatieregels. Het algoritme kijkt immers naar de support van een itemset, niet van één woord op zichzelf.

Ook werd een aantal overlappende woordgroepen aangemaakt. De woorden ‘restaurant’ en ‘fruit’ hadden bijvoorbeeld op zich zelf al voldoende support, maar vallen ook onder de woordgroep ‘eten’.

Alle in deze paragraaf toegepaste handelingen leiden tot het resultaat van 934 woordgroepen.

10.5 Toepassen datamining en kwaliteitsmetingen

De CBA tool werd gebruikt om het CBA-RG algoritme toe te passen op de dataset. De regels zijn gegenereerd met een minimale support van één procent en een minimale confidence van tachtig procent. De tool is ook aangegeven niet meer dan 80.000 regels te genereren, alleen regels te genereren uit itemsets die niet groter zijn dan tien items.

CBA leverde de associatieregels in één document en de itemsets in een ander document. Omdat deze informatie zich in aparte documenten bevond, is de support van de consequent handmatig toegevoegd in het document met de associatieregels.

Group	Measure
1	Odds Ratio Yule's Q Yule's Y
2	Cosine Jaccard
3	Support Laplace
4	ϕ-coefficient Collective Strength Piatetsky-Shapiro's
5	Gini Index Goodman-Kruskal's
6	Interest Factor Added Value Klosgen
7	Mutual Information Certainty Factor Kappa

Figuur 17: gegroepeerde kwaliteitsmaten [O82]

Tan et al [O81] geven aan dat kwaliteitsmaten soms tegenstrijdige waarden toekennen aan associatieregels. Dr. T.M. Heskes, expert op het gebied van datamining aan de Radboud Universiteit, stelt dat er voor geen enkele kwaliteitsmeting op dit gebied standaard harde drempelwaarden zijn. De keuze voor de grenswaarde is om deze reden steekproefsgewijs verlopen, wanneer een grenswaarde een goede scheiding leek aan te brengen is deze gebruikt.

Op basis van eigenschappen van de kwaliteitsmaten, hebben Tan et al. [O82] kwaliteitsmaten gegroepeerd. In figuur 17 staat de groepering zoals Tan et al deze hebben aangebracht. Uit elk van deze groepen is één kwaliteitsmaat gekozen om te gebruiken op de dataset. Op deze manier is getracht gevolgen van de keuze van een verkeerde kwaliteitsmaat te voorkomen. Wanneer een verkeerde maat gekozen zou worden, zouden juist de slechte regels overblijven en de goede verwijderd worden. Door met verschillende uiteenlopende kwaliteitsmaten, elke keer een relatief klein aantal regels te verwijderen, wordt het risico per kwaliteitsmaat beperkt, en wordt hoogstens een klein deel van de beste regels verwijderd.

Naast de kwaliteitsmaten die gekozen zijn aan de hand van de groepering van Tan et al. [O82] is de confidence gebruikt. Deze kwaliteitsmaat is gebruikt omdat hij onderdeel vormt van het CBA-RG algoritme. De kwaliteits maat 'Lift' is gebruikt in plaats van de interest factor. Aangezien de Lift afgeleid is van de kwaliteitsmaat interest factor, kunnen we deze maat binnen de groepering van Tan et al. plaatsen.

De volgende kwaliteitsmetingen zijn binnen dit onderzoek gebruikt:

Groep Tan et al [O82]	Kwaliteitsmaat	Berekening
3	Support	f11 / N
	Confidence	f11 / f10
6	Lift	(f11 / f10) / (f10 / N)
2	Jaccard	f11 / (f1+ + f+1 – f11)
7	Certainty Factor	((f11 / f1+) – (f+1 / N)) / (1 – (f+1 / N))
5	Gini Index	(f1+ / N) * ((f11 / f1+)² * (f10 / f1+)²) – (f+1 / N) ² + (f0+ / N) * ((f01 / f0+)² * (f00 / f0+)²) – (f+0 / N) ²
1	Odds Ratio	(f11 * f00) / (f10 * f01)
4	Correlation-coëfficiënt	(N * f11 - f1+ * f+1) / (√ (f1+ * f+1 * f0+ * f+0))

Tabel 13: Binnen dit onderzoek gebruikte kwaliteitsmaten.

De kwaliteitsmaten support en confidence worden gebruikt door het datamining programma CBA. CBA-RG word opgedragen alleen regels op te leveren wanneer zij een minimale support hebben van één procent en een minimale confidence van tachtig procent. De support en de confidence van de regels worden door het programma uitgerekend en bij de resultaten vermeld.

De overige kwaliteitsmaten worden daarna handmatig toegepast. Voor de handmatig toegepaste kwaliteitsmaten geld dat de grens wordt bepaald aan de hand van de regels. Dit betekend dat een optionele grenswaarde wordt gekozen (waarbij niet meer dan een kwart van de regels verloren gaat), vervolgens wordt gekeken naar de regels rond de grens. Wanneer zou blijken dat een groot gedeelte van de regels die net buiten de grens vallen erg nuttig en waardevol zijn, wordt een andere optionele grens gekozen. Voor deze grens wordt dezelfde werkwijze toegepast.

10.6 Selecteren nuttige associatieregels

Voor het selecteren van nuttige associatieregels wordt vier doctorandussen met expertise op het gebied van marketing gevraagd de mogelijke consequenten te beoordelen. Deze doctorandussen doceren allen aan dezelfde HEAO en kennen elkaar.

De doctorandussen werd gevraagd het nut voor adverteerders van de kennis over het voorkomen van deze woorden of eigenschappen te beoordelen voor alle mogelijke consequenten. De beoordeling geschied d.m.v. een schaal van 5. Zij kunnen de woorden en eigenschappen een 1,2,3,4 of 5 toewijzen, waar de 1 staat voor totaal niet nuttig en de 5 voor erg nuttig. Van elke consequent wordt het gemiddelde van de beoordeling van de doctorandussen aangehouden. Vervolgens wordt voor alle regels aan de hand van hun consequent bekeken wat voor een beoordeling bij de regel hoort. Voor de regels die twee consequenten hebben, wordt de hoogste beoordeling genomen.

11 Resultaten experiment Associatie analyse

De resultaten van het experiment waarvan de aanpak beschreven is in hoofdstuk 10 zijn in dit hoofdstuk uiteengezet. Een analyse van de onderzoeksobjecten, de gebruikte kwaliteitsmaten en het selecteren van nuttige associatie regels komen aan bod. Vervolgens worden de resultaten geanalyseerd en wordt de kwaliteit van de resultaten beoordeeld.

11.1 Onderzoeksobjecten

De onderzoeksobjecten binnen dit experiment vormen de profielen van gebruikers op twee sociale netwerksites. Van de dataset is 4,9 procent afkomstig van girlsonly. De overige ruim 95 procent is afkomstig van Ikku.nl.

De volgende informatie is gebruikt dit experiment:

- Site {Ikku.nl, Girlsonly.nl}
- Geboortejaar
- Provincie {Gelderland, Limburg, Friesland, Utrecht, Zuid-Holland, Noord-Holland, Overig, Overijsel, Noord-Braband, Groningen, Drente, Zeeland, Flevoland}
- Opleiding {vwo, havo, mavo, vmbo-t, vmbo-g, vmbo-k, vmbo-b, vbo, anders}
- Geslacht {man, vrouw}
- Geaardheid {hetero, homo, biseksueel}
- Relatie {ja, nee}
- Huidskleur {blank, bruin, zwart}
- Kleur ogen {bruin, blauw}
- 934 woordgroepen uit de losse tekst

De woordgroepen uit de losse tekst, zijn samengesteld aan de hand van selectiecriteria in paragraaf 10.4.3.

11.2 Gebruik van kwaliteitsmaten

11.2.1 Support en confidence

CBA-RG leverde 23.510 associatieregels op, die voldeden aan de minimum support en confidence.

11.2.2 Lift

Van de 23.510 associatieregels, bleken de antecedent en consequent bij 22 regels negatief te correleren, de laagste liftwaarde was 0,9256. Er zijn geen regels gevonden waar antecedent en consequent geheel onafhankelijk (liftwaarde exact één) bevonden werden. Bij de overige associatieregels correleren antecedent en consequent positief.

Besloten is een grenswaarde te kiezen van 1,5. Alle associatieregels met een lagere liftwaarde (4380 regels) worden verwijderd, waardoor het totale aantal regels op 19.130 uitkomt. Opvallend is dat er slechts twee associatieregels een liftwaarde tussen de 1,5 en de 5 hebben. Er zijn 13 regels met een liftwaarde boven de 85. De meest positieve correlatie die gevonden is, had een liftwaarde van 2.599.

11.2.3 Jaccard

De regels met een Jaccard waarde onder de 0,10 zijn verwijderd. Dit leidde tot de verwijdering van 4623 regels. Veel van deze regels hadden dezelfde consequent. Dit geeft aan dat het minder nuttige regels zijn. De woordgroep ‘eten’ was de meest voorkomende consequent, deze kwam namelijk in ruim de helft van de verwijderde regels als consequent voor. Na het verwijderen van regels aan de hand van de jaccard, bleven er nog 14.508 regels over.

11.2.4 Certainty Factor

De certainty factor [O83] ligt binnen de nog aanwezige regels tussen de 0,765 en 1. De regels met een certainty factor onder de 0,80 verwijdert. Dit zijn 1.484 regels, het aantal regels komt daardoor uit op 13.024.

11.2.5 Gini Index

Bij de Gini index liggen de waarden van de associatieregels tussen de -0,0297 en de 0,1473. De grenswaarde is op 0,003 gelegd, dit betekend dat alle associatieregels met een lagere waarde verwijderd zijn. Het gaat hierbij om 2.145 regels, deze verwijdering heeft als resultaat dat het nieuwe totale aantal regels op 10.880 ligt.

11.2.6 Odds Ratio

De odds ratio is gebruikt om het aantal regels verder te beperken. De odds ratio geeft op de associatieregels waarden tussen de 4,1604 en 3735,0333. Er zijn slechts achttien regels tussen de 1.000 en 2.000 en één regel met een waarde boven de 2.000 (namelijk 3.735,033). Daarnaast zijn er 89 regels waarvan de waarde niet berekend kan worden.

Alle associatieregels met een waarde onder de 5,5 zijn verwijderd. Dit betekend dat er 1233 regels verwijderd worden. Op het eerste gezicht lijken hier een aantal interessante regels tussen te zitten, maar deze blijken afkomstig te zijn van woorden die gezamenlijk in veelgebruikte vragenlijsten voorkomen. Wat er wordt verstaan onder deze vragenlijsten, is te lezen in paragraaf 7.3.3.

De regels waarover de odds ratio niet goed berekend kon worden, zijn grotendeels regels die door groepering ontstaan zijn. Van de 89 regels die een fout bij de berekening gaven omdat er niet gedeeld kan worden door nul, bleken 83 regels te zijn veroorzaakt door groepering. Bij deze berekeningen moest er gedeeld worden door nul omdat er geen transacties waren waar wel de antecedent in zat, maar niet de consequent. Deze associatieregels hadden immers een confidence van honderd procent.

Drieëntachtig regels bleken veroorzaakt te zijn door een overlappende groep bestaand uit meerdere woordgroepen, zoals beschreven in paragraaf 10.4.5. Deze regels werden verwijderd. Het is mogelijk dat er regels zijn met een lagere confidence die ook door deze manier van groeperen ontstaan zijn, deze zijn echter lastiger op te sporen.

Bovenstaande verwijderingen hebben ertoe geleid dat er nog 9.564 associatieregels zijn.

11.2.7 Correlation Coëfficiënt

Bij de correlation coefficient liggen de waarden van de associatieregels tussen de 3692,33 en de 133095,0147. De grenswaarde is op 4.500 gelegd, alle associatieregels met een lagere waarde zijn verwijderd. Dit zijn 1.713 regels, wat als resultaat heeft dat het nieuwe totale aantal regels op 7.851 ligt.

11.3 Selecteren nuttige associatieregels

Voor het selecteren van nuttige associatieregels is vier doctorandussen met expertise op het gebied van marketing gevraagd de mogelijke consequenten te beoordelen door hen een 1,2,3,4 of 5 toewijzen, waar de 1 staat voor totaal niet nuttig en de 5 voor erg nuttig. Wanneer een drie als beoordeling wordt gegeven is dit dus een neutraal standpunt ten op zichtte van de eigenschap of het woord.

De gemiddelde beoordeling van de vier personen over de eigenschappen van een profieleigenaar is 2,89, waar de gemiddeldes per persoon tussen de 2,50 en 3,28 liggen. De gemiddelde beoordeling van de gebruikte woordgroepen in het profiel ligt op 2,79, met de gemiddeldes per persoon tussen de 2,43 en 3,17.

Door de associatieregels en de beoordeling in een tabel te zetten is dmv SQL de beoordeling van iedere regel berekend. Voor de regels die twee consequenten hebben, is de hoogste beoordeling genomen.

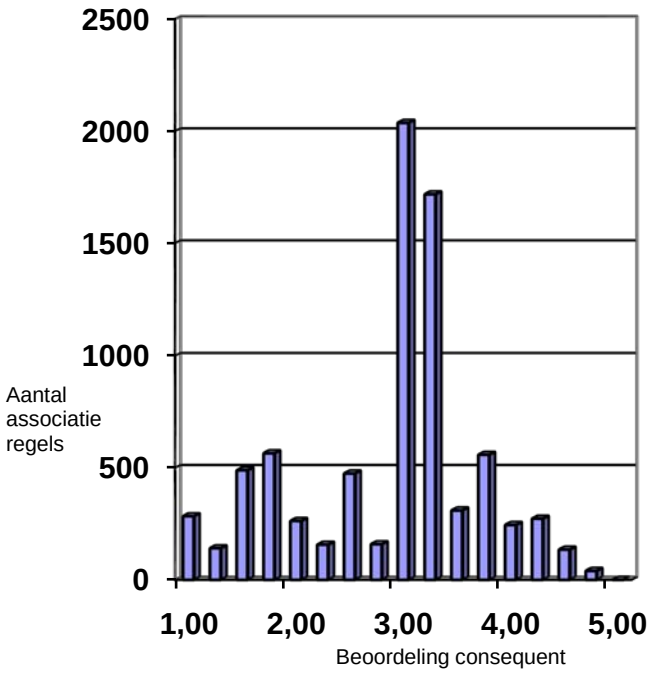


Fig 18: Beoordeling van consequenten

Beoordeling consequent	Aantal associatieregels
1,00	284
1,25	141
1,50	489
1,75	565
2,00	263
2,25	157
2,50	475
2,75	159
3,00	2037
3,25	1718
3,50	310
3,75	558
4,00	245
4,25	273
4,50	135
4,75	40
5,00	0

Tabel 14: Beoordeling van consequenten

consequent	Aantal regels
verliefd	16
vakantie	40
relatie	1
pinpas	4
ijs	104
film	112
bioscoop	106
auto	35
alcohol	29
totaal	447

Tabel15: Aantal regels per consequent, beoordeling > 4

Hoewel alle regels met een score boven de 3,00 interessant zijn, is dit nog steeds een erg groot aantal associatieregels, namelijk 3.279 regels. Daarom worden in bovenstaande tabel alleen de regels met als minimale score een 4 wordt gegeven. Een nadeel van de selectie op de score, is dat er slechts 18 consequenten zijn met een score boven de 4. Tien hiervan komen enkel voor in regels die niet hebben voldaan aan de kwaliteitsmaten. In tabel 23 is voor elke consequent één regel weergegeven.

Opvallend in de selectie is, dat alle regels die het woord ‘vakantie’ of ‘alcohol’ als consequent bevatten, twee antecedenten hebben. Regels die ‘vakantie’ als consequent hebben, bevatten allemaal een van de volgende antecedenten: ‘ijs’, ‘school’, ‘eten’. Regels die ‘alcohol’ als consequent hebben, bevatten allemaal een van de volgende antecedenten: ‘ijs’, ‘hemel’, ‘school’, ‘eten’. Die overeenkomst kan er op wijzen dat de woordgroepen ‘vakantie’, ‘alcohol’, ‘ijs’, ‘school’ en ‘eten’ samen in een vragenlijst voorkomen.

Ook valt op dat de consequenten ‘film’ en ‘bioscoop’ veel antecedenten gemeen hebben. Dit is niet verbazend omdat men het al snel over een film heeft wanneer het woord ‘bioscoop’ in een tekst genoemd wordt.

Regel	antecedent 1	antecedent 2	->	consequent 1	confidence	support	Beoordeling
22119	pinpas		->	verliefd	90,19	1,306	4,25
15402	mail	ijs	->	vakantie	92,63	1,554	4,75
19175	verwijderd	ruzie	->	relatie	96,31	1,373	4,25
16739	hemel	branden	->	ijs	97,54	1,239	4,25
21914	provider		->	film	99,65	3,69	4,25
14297	uitdaging	eten	->	bioscoop	98,36	2,81	4,5
22820	steel		->	auto	89,18	2,46	4,25
15796	uiterlijk	ijs	->	alcohol	95,64	1,278	4,5

Tabel 16: Aantal voorbeeld regels met een beoordeling boven de 4

11.4 Analyse van Resultaten

In deze paragraaf worden de resultaten die in dit hoofdstuk gepresenteerd zijn geanalyseerd.

11.4.1 Invloed van vragenlijsten

Na een korte inventarisatie van een aantal opmerkelijke associatieregels, bleek dat deze vaak gezamenlijke aanwezigheid van woorden in een vragenlijst hadden die aangeboden werd op ikku.nl. Het gaat hier om de vragenlijsten die veel jongeren zelf op hun profiel plaatsen, deze vragenlijsten zijn verder beschreven in paragraaf 7.3.3.

Door middel van SQL-queries op de dataset en de door Ikku.nl aangeboden vragenlijsten, werd bekeken of associatieregels dankzij de vragenlijsten ontstaan zijn door of een verhoogde support en confidence hebben bereikt.

Een aantal associatieregels werd door meerdere vragenlijsten ondersteund. Alle vragenlijsten dragen namelijk gezamenlijk 78.573 maal bij aan de support en confidence van de regels. Echter het gaat om 7.756 associatie regels waarvan het aantal voorkomens waarschijnlijk is beïnvloed door het gebruik van de vragenlijsten.

Dit zijn bijna alle associatieregels. Dat dit een groot deel van de associatieregels omvat, kan toegeschreven worden aan meerdere redenen, bijvoorbeeld de onderwerpen van de vragenlijsten en het in het dagelijks leven ook vaak voorkomen van deze woorden.

Zo zijn de vragenlijsten gericht op jongeren en bevatten ze daarom voor hen interessante onderwerpen. Dit verklaard de grote overlap in woordgebruik binnen de vragenlijsten en op het overige deel van de profielen. Wanneer er een grote overlap in woordgebruik is, wordt het waarschijnlijker dat een grote overlap in associatieregels ontstaat.

De aanwezigheid van woorden op vragenlijsten had mogelijk beter in een eerder stadium onderzocht kunnen worden. Echter, de gezamenlijke aanwezigheid van woorden in een vragenlijst, is vaak ook de reden voor het gezamenlijk voorkomen van deze woorden in het algemeen. Dit komt omdat het merendeel van de vragenlijsten zich focust op één onderwerp. Wanneer ‘school’ en ‘vak’ vaak samen voorkomen, kan dit niet alleen toegewezen worden aan de vragenlijsten. Deze woorden worden immers door jongeren vaak samen gebruikt.

De omvang van de invloed van de vragenlijsten kan moeilijk bepaald worden. Een reden hiervoor is dat de vragenlijsten als tekst in het profiel zijn geplaatst, tussen de andere losse tekst. Wanneer vragenlijsten in een apart formulier op het profiel zou worden geplaatst (en opgeslagen in de database) zou een vragenlijst meer herkenbaar zijn. Ook kan men dan bekijken of per vragenlijst verschilt hoe vaak deze is ingevuld, dit heeft namelijk een sterk aandeel in de invloed van de vragenlijst op de gehele tekst en in dit geval de associatieregels die daar van zijn afgeleid.

11.5 Kwaliteit resultaten

Van de woorden uit woordenlijst die in het associatie analyse experiment is gebruikt, is de context niet bekend. Wanneer we bijvoorbeeld kijken naar het woord ‘hond’ is niet duidelijk of deze persoon veel van honden houd, zelf een hond bezit, een hekel heeft aan honden of bang is voor honden. (zie aanbevelingen)

11.6 Validiteit

Een van de vraagstukken binnen dit onderzoek is te beoordelen of de gegenereerde associatie regels ook bruikbaar zijn voor adverteerders. Volgens Adomavicius en Tuzhilin [O43] is het gebruikelijk een domein expert de gevonden regels te laten valideren als vorm van postanalyse. Voor dit onderzoek is een aantal economiedocenten gevraagd de gegenereerde regels te valideren.

12 Combinatie databetrouwbaarheid & associatie analyse

In dit hoofdstuk worden de resultaten van het databetrouwbaarheidsonderzoek en het associatie analyse experiment naast elkaar gelegd. Twee raakvlakken van de twee experimenten komen in dit hoofdstuk aan bod, namelijk:

- Databetrouwbaarheid van associatieregels.

Bekeken wordt wat er gezegd kan worden over de databetrouwbaarheid van de associatieregels, nu we onderzoek hebben gedaan naar de databetrouwbaarheid van de profielen.

- Geschatte aantal profielen met vragenlijst als input voor associatie analyse.

Uit het associatie analyse experiment is gebleken dat standaard vragenlijsten op de profielen invloed hebben op de resultaten van het associatie analyse experiment. Daarom wordt aan de hand van antwoorden uit het databetrouwbaarheids onderzoek ingeschat hoeveel van de profielen (welke dienden als input van het associatie analyse regelement) een standaard vragenlijst bevatten.

12.1 Databetrouwbaarheid van associatieregels

Aan de hand van het databetrouwbaarheidsonderzoek is nu bekend over welke onderwerpen veel gelogen wordt. Het associatie analyse experiment heeft uitgewezen wat de meest gebruikte woorden en gegevens zijn en welke verbanden hier tussen getrokken kunnen worden. Volgens het databetrouwbaarheidsonderzoek staat de volgende informatie niet in alle gevallen correct op het profiel:

- Woonplaats
- Opleiding
- Relatiestatus
- Leeftijd
- Geaardheid
- Uiterlijk
- Oogkleur
- Hobby's
- Huidskleur
- Provincie
- Sterrenbeeld
- Geslacht
- Naam
- Achternaam
- Geboortedatum
- Idolen
- Favoriete films
- Favoriete muziek
- Huisdieren

Geen van deze onderwerpen komt voor in de resulterende associatieregels. Bij een aantal onderwerpen is dit ook bijna niet mogelijk. Woonplaats bijvoorbeeld is om privacy overwegingen buiten het datamining experiment gelaten. Wanneer het wel gebruikt was zijn zoveel uiteenlopende antwoorden, dat de kans klein is dat dit terug te vinden is in associatieregels. 'Amsterdam' viel (vanuit de losse tekst) wel binnen de frequentiegrens, maar is verderop in het selectieproces alsnog afgevallen. Men kan zich voorstellen dat bij geboortedata, favoriete films, e.d. hetzelfde zich voor zal doen. Bij onderwerpen als uiterlijk, huidskleur e.d. is er vaak niet duidelijk over welke persoon er geschreven wordt. Binnen het databetrouwbaarheidsonderzoek heeft de focus alleen op de ondervraagde zelf gelegen.

In hoeverre er oncorrecte data voorkomt in de associatieregels is dus niet precies te zeggen. Echter wel kan geconstateerd worden dat maar een klein aantal gegevens uit associatieregels die te plaatsen is, binnen de onderwerpen waar mensen over hebben toegegeven dat data op hun profiel niet klopt. Nu zijn niet alle gegevens uit de associatieregels in een verplichte gesloten vraag in de enquête aan de orde gekomen. Maar omdat mensen deze gegevens niet bij de open vraag hebben aangegeven, kunnen we hier voorzichtig uit concluderen, dat de gegevens in de associatieregels over het algemeen correct op het profiel van de gebruikers van deze sites staan.

12.2 Geschatte aantal profielen met vragenlijst als input voor associatie analyse experiment

In Tabel 12 uit paragraaf 8.6 staat genoteerd dat het gemiddelde aantal ondervraagden met een vragenlijst op 13,8 ligt. Echter uit dezelfde tabel is af te lezen dat het aantal ondervraagden met een vragenlijst per sociale netwerksite veel verschilt.

Wanneer we aannemen dat de respondenten van het databetrouwbaarheidsonderzoek representatief zijn voor de eigenaren van de profielen (bij dezelfde site) die tot de dataset behoren, kunnen we aan de hand daarvan inschatten hoeveel mensen een vragenlijst op hun profiel hebben staan. Omdat de verdeling van onderzoeksobjecten onder de sites bij dit experiment niet het zelfde is als binnen het databetrouwbaarheid onderzoek, worden de cijfers daarop gecorrigeerd. Het aantal profielen met een vragenlijst ingeschat door het percentage leden met een vragenlijst afkomstig uit het databetrouwbaarheidsonderzoek, vermenigvuldigd wordt met het aantal leden per sociale netwerksite. De in en output van deze berekening staat weergegeven in tabel 17. Uit de schatting blijkt dat bijna de helft van de profielen minstens één vragenlijst zou moeten bevatten.

Site	Aantal met vragenlijst in data-betrouwbaarheid onderzoek	Verdeling profielen binnen datamining experiment	Aantal profielen binnen datamining experiment	Geschatte aantal profielen met vragenlijst
Girlsonly	3,2 procent	4,9 procent	3.457	110
Ikku.nl	51,7 procent	95,1 procent	66.898	34.591
Gurlz.nl	40,9 procent	0 procent	0	0
Hunkz.nl	20,0 procent	0 procent	0	0
Totaal	13,8 procent	494	70.355	34602

Tabel 17: Aantal respondenten met vragenlijst per site

Er is binnen het databetrouwbaarheidsonderzoek geen vraag gesteld over het aantal vragenlijsten dat men op het profiel had staan. Ook is niet duidelijk hoe gebruikersaantallen verdeeld zijn over de vragenlijsten. Om aan de hand van de bekende gegevens toch berekeningen te kunnen maken doen we de volgende aannames:

- Wanneer minstens één vragenlijst op een profiel staat, is dit ook altijd één vragenlijst en staan er in geen enkel geval meerdere vragenlijsten op een profiel.
- Er is niet gelogen op het bezit van een vragenlijst
- Het aantal vragenlijsten op de data van het verkrijgen van de datasets, is gelijk aan het aantal vragenlijsten op het moment van het databetrouwbaarheidsonderzoek.
- Iedere vragenlijst op een gelijk aantal profielen geplaatst is ten opzichte van andere vragenlijsten.
- De onderzoeksobjecten binnen het databetrouwbaarheidsonderzoek zijn representatief voor de eigenaren van de profielen op de door hen gebruikte sociale netwerksite.
- Er worden geen andere vragenlijsten gebruikt dan de vragenlijsten op Ikku.nl

Dat de laatste aanname niet klopt, is bekend. Echter we nemen dit toch aan omdat niet alle bestaande vragenlijsten onderzocht kunnen worden. Voor enkele andere aannamen zijn er aanwijzingen dat zij niet volledig correct zijn, bijvoorbeeld de aanname dat elke vragenlijst op een gelijk aantal profielen is geplaatst. Echter deze aannamen maken het mogelijk een schatting te maken van het aantal profielen met vragenlijst.

Bovenstaande aannames zouden er toe leiden dat wanneer 346.022 profielen een vragenlijst bevatten, alle 22 vragenlijsten een gelijk aantal verschijningen op profielen heeft. Dit zou per profiel op 15.728,3 voorkomens per vragenlijst uit komen. Dit betekend een support van 22 procent. Wanneer bovenstaande aannames zouden kloppen, hebben alle woorden uit de vragenlijsten dus een behoorlijke invloed op de associatieregels.

Dat de aannames waarschijnlijk niet volledig kloppen blijkt uit het feit dat de associatieregels met een support boven de 22 procent (van de associatieregels opgeleverd door CBA), allemaal regels zijn die niet afkomstig zijn uit de losse tekst. Dergelijke regels zijn door middel van analyse met kwaliteitsmaten verwijderd. Regels die wel een support boven de 22 procent uitkomen zijn bijvoorbeeld, Als iemand blauwe ogen heeft, is het waarschijnlijk dat hij/zij ook een blanke huid heeft.

De hierboven gestelde aannames blijken dus niet allen te kloppen. Ik verwacht dat het aantal profielen waarop een vragenlijst staat ten opzichte van andere vragenlijsten behoorlijk zal variëren. Wel heeft deze analyse een indruk gegeven van het effect van vragenlijsten op het samen voorkomen van woorden in een profiel.

13 Discussie

In dit hoofdstuk komen punten aan bod welke anders aangepakt hadden kunnen worden, maar wegens onder andere praktische redenen en afbakening niet verder onderzocht zijn.

13.1 Sociale netwerksites

De hoeveelheid wetenschappelijke artikelen over sociale netwerksites is relatief klein, omdat dit nog een jong onderzoeksonderwerp is en een beperkt aantal wetenschappers onderzoek doet op dit vakgebied. Dit heeft als gevolg dat er in hoofdstuk 5 en 6 veel gerefereerd naar artikelen van een beperkt aantal onderzoekers en er veel gerefereerd is naar onderzoek onder gebruikers van sociale netwerksites in de Verenigde Staten. Wanneer wetenschappelijke artikelen ontbraken zijn soms persberichten of krantenberichten als bron gebruikt. Vanzelf sprekend is gekozen voor zo betrouwbaar mogelijke bronnen.

13.2 Databetrouwbaarheidsonderzoek

De sociale wenselijkheidsvragen die terugkomen in de enquête zijn gevalideerd onder Amerikaanse kinderen in de leeftijd van acht tot achttien jaar. De jongeren waarbij de enquête is afgenomen hebben de Nederlandse nationaliteit. Daarnaast behoord een deel van deze jongeren niet tot de leeftijdscategorie waaronder gevalideerd is. De testgroep sluit niet volledig aan op de validatiegroep, waardoor culturele verschillen en leeftijdsverschil een rol kunnen spelen in de beantwoording van de vragen.

De vragenlijst die gebruikt is voor het databetrouwbaarheidsonderzoek, is voor gebruik getest door een groep van vijfendertig mensen. De test onder de groep van 35 mensen kan niet voldoende generaliseert worden om als goede validatie van de vragenlijst te dienen. De groep waarbij deze test is afgenomen, verschilt op een aantal punten met de onderzoeksgroep van het databetrouwbaarheidsonderzoek. Verschillen zijn met name te vinden in leeftijd, opleidingsniveau en geslacht. Het is mogelijk dat de testgroep en onderzoeksgroep een andere interpretatie of mate van begrip van de vragenlijst hebben. Om praktische redenen is gekozen de test te houden onder een (voor dit onderzoek) niet ideale testgroep. De test heeft zijn nut desondanks wel bewezen. De test bracht een aantal onduidelijkheden rondom vragen naar voren. Aan de hand van deze bevindingen is de vragenlijst verbeterd.

13.3 Sociale wenselijkheid

De CSDTC is gevalideerd in een leeftijdscategorie tot en met achttien jaar oud. Carifio heeft zijn verkorte vragenlijst gevalideerd onder Amerikaanse kinderen van tien tot en met vijftien jaar oud. Een deel van mijn onderzoeksobjecten behoord ook tot deze leeftijdscategorie; de rest heeft een leeftijd net buiten deze categorie. Er kan niet met zekerheid gezegd worden dat vragenlijst B de sociale wenselijkheid ook goed test onder kinderen buiten de leeftijdscategorie.

Het zou ook mogelijk zijn geweest zijn afhankelijk van de leeftijd van het onderzoeksobject te kiezen voor een vragenlijst, waardoor er binnen het gehele onderzoek twee of meerdere vragenlijsten gehanteerd worden. De zwakke punten in de verschillende vragenlijsten kunnen dan van elkaar verschillen, waardoor resultaten minder vergelijkbaar is. Daarom is niet gekozen voor een leeftijdsafhankelijk gebruik van vragenlijsten.

Een vragenlijst die gevalideerd was op alle leeftijden van de onderzoeksobjecten was niet voor handen. Veel sociale wenselijkheid vragenlijsten zijn voor volwassenen of kinderen. Een dergelijke vragenlijst voor jongeren heb ik niet kunnen vinden. Omdat de gemiddelde leeftijd binnen een validatie op volwassenen ver af ligt van de gemiddelde leeftijd van de onderzoeksobjecten, kon beter gekozen worden voor een vragenlijst gevalideerd onder kinderen.

Hier wordt echter de kanttekening geplaatst, dat dit niet het enige kenmerk is van een validatie waar op gelet dient te worden: ook etnische achtergrond, nationaliteit, religie, welvaart zijn onder andere punten

waarop de validiteit geoordeeld kan worden.

13.4 Associatie analyse experiment

Niet alle associatie analyse algoritmes zijn bekeken voor gekozen is welk algoritme te gebruiken. De voornaamste reden voor de keuze van het Apriori algoritme is dat dit een van de weinige algoritmes is welke zich bewezen heeft en veelvuldig in wetenschappelijk onderzoek gebruikt is. Dit maakt het makkelijker kwaliteitsmetingen en resultaten van dit experiment te vergelijken met andere experimenten.

De gebruikte dataset voor het associatie analyse experiment is grotendeels afkomstig van Ikku.nl, omdat ikku.nl meer profielen bevat dan Girlsonly. Dit kan een vertekend beeld geven.

Beheerders van sociale netwerksites hebben er baat bij profielen die niet meer gebruikt worden te behouden. Weliswaar zou het goed zijn de bestanden op te schonen, meer baat hebben zij bij hoge ledenaantallen. Een hoog ledenaantal trekt meer publiciteit en adverteerders, wat de sites helpt verder te groeien. Zelfs het verwijderen van je eigen profiel is op veel sites niet mogelijk, dan wel erg lastig. Omdat sites het een gebruiker moeilijk maken zijn profiel te verwijderen en zij de beheerders zelf meestal geen oude profielen verwijderen, zal een gedeelte van de gepubliceerde gebruikersaantallen niet meer actief zijn op de betreffende sociale netwerksite. Ook worden er in de ledenaantallen mensen dubbel geteld omdat één persoon meerdere identiteiten kan hebben op een sociale netwerksite. Dat mensen daadwerkelijk meerdere identiteiten bezitten op dezelfde site, is ook terug te vinden in paragraaf 6.6 en 8.3.

Vier doctorandussen hebben de consequenten beoordeeld zoals staat beschreven in hoofdstuk 10.6. Dit is een kleine steekproef. Met name daarom, kan het gemiddelde van deze beoordelingen niet als representatief worden gezien voor de algemeen heersende mening binnen het economische vakgebied. Dat dit gemiddelde toch gebruikt is om nuttige associatieregels te selecteren, maakt de marketingwaarde van de selectie minder bewezen.

De doctorandussen die de consequenten beoordeelden zijn collega's van elkaar. Dat zij elkaar kennen kan er toe leiden dat zij met elkaar overlegd hebben. De docenten is gevraagd dit niet te doen en de beoordelingen verschillen dusdanig dat dit niet waarschijnlijk is. Ook hebben zij weinig belang bij het geven van een gelijke beoordeling, aangezien de beoordeling van de consequenten per doctorandus niet openbaar worden gemaakt. Ook is het mogelijk dat hun antwoorden dichter bij elkaar liggen dan dat het geval zou zijn wanneer doctorandussen gevraagd worden die elkaar niet kennen, omdat zij mogelijk al eens met elkaar gesproken hebben over de scheiding tussen nuttige en niet nuttige informatie voor adverteerders.

14 Aanbevelingen

Bevindingen opgedaan binnen dit onderzoek, geven interessante input voor vervolgonderzoek. In dit hoofdstuk staan aanbevelingen voor vervolgonderzoek. Natuurlijk kunnen resultaten uit dit onderzoek als uitgangspunt dienen voor vervolg onderzoek, door bijvoorbeeld te onderzoeken of de databetrouwbaarheid verschilt per sociale netwerksite. Maar men kan ook andere uitgangspunten gebruiken om de resultaten breder te valideren.

14.1 Databetrouwbaarheid

Binnen het databetrouwbaarheidsonderzoek is gekozen voor een sociale wenselijkheidsvragenlijst, welke weergegeven is in figuur 6. Zoals ook in hoofdstuk 14 wordt beschreven, sluit de vragenlijst die gebruikt is niet perfect aan op de onderzoeksgroep. Er zou bekeken kunnen worden of gebruik van een andere sociale wenselijkheidsvragenlijst andere resultaten oplevert.

De resultaten uit dit onderzoek zijn niet per definitie van toepassing op andere sociale netwerksites. Verschillen tussen sociale netwerksites kunnen leiden tot verschillen in databetrouwbaarheid. Eigenschappen van sociale netwerksites, die verschillen in databetrouwbaarheid zouden kunnen veroorzaken zijn o.a.:

- eigenschappen van de gebruikersgroep (bijvoorbeeld culturele achtergrond of leeftijdscategorie)
- de heersende norm op de site
- eigenschappen van invoervelden (of zij een standaardwaarde bevatten of verplicht zijn)
- regels op de site (bijvoorbeeld leeftijdsrestricties)

Meer informatie over redenen voor verminderde databetrouwbaarheid staat in hoofdstuk 6. Voor een aantal van deze eigenschappen van sociale netwerksites zou men de resultaten kunnen corrigeren, om deze vergelijkbaar te maken met resultaten uit dit onderzoek.

In welke mate de eigenschappen van sociale netwerksites invloed hebben op de databetrouwbaarheid kan in verder onderzoek aan bod komen. Mogelijk kan men op die manier tot een model komen waarin aan de hand van eigenschappen van de site en zijn doelgroep geschat kan worden hoe betrouwbaar profieldata is.

Ook kan men (wanneer er meer bekend is over databetrouwbaarheid van profielen op sociale netwerksites) proberen de data te corrigeren aan de hand van associatieve of causale regels welke leiden tot meer databetrouwbaarheid. Hierbij is van belang de profieldata te kunnen verifiëren om te onderzoeken of de correctie geslaagd is geweest.

14.2 Experiment Datamining

Zoals in hoofdstuk 9 besproken, is de context niet bekend van de woorden uit woordenlijst die gebruikt is in het associatie analyse experiment. De semantische oriëntatie van de woorden is daardoor niet duidelijk. Omdat niet bekend is of een persoon positief of negatief over een bepaald onderwerp schrijft, is het effect van gebruik van deze informatie te gebruiken voor een advertentie beperkt voorspelbaar. Door de jaren heen zijn er diverse methodes [O84], [O85], [O86] gevonden om de semantische oriëntatie van woorden automatisch te beoordelen. Deze methodes zouden gebruikt kunnen worden om de waarde van de resultaten van de associatie analyse voor het gebruik in advertenties te verhogen.

Associatieve regels uit ander wetenschappelijk onderzoek op een vergelijkbare doelgroep kunnen gebruikt worden als aanvulling op de resultaten uit dit onderzoek. Door dit te doen worden de associatieregels resulterende uit dit onderzoek uitgebreid met extra antecedenten en/of consequenten. Dit kan de mogelijke toepassing van de resultaten uit dit onderzoek flink vergroten. Ter illustratie: wanneer we associatieregel 21914 pakken uit tabel 16 en uit eerder onderzoek blijkt dat voor 90 procent van de mensen die regelmatig over films communiceren ook veel over games communiceren, dan zouden die twee regels gecombineerd kunnen worden. Dit zou dan het volgende resultaat geven: provider > film > games.

In het experiment waarin associatie analyse is toegepast, is het programma CBA gebruikt. CBA biedt de mogelijkheid met minimale item support te werken. Dit houdt in dat er per item een minimale support gesteld kan worden, in plaats van het instellen van één supportwaarde voor alle items [O53]. In het datamining experiment beschreven in hoofdstuk 9 en 10 is gewerkt met één support waarde. Hierdoor komen zeldzame items (bijvoorbeeld één van de woorden uit de losse tekst) minder snel voor in regels en veelvoorkomende items (bijvoorbeeld van een verplicht meerkeuzeveld) komen juist in veel regels voor. In dit onderzoek is de aanname gedaan dat associatieregels alleen interessant zijn voor het plaatsen van advertenties, wanneer deze gedragen worden door een redelijk groot aantal profielen. Uit verder onderzoek zal moeten blijken of het gebruik van een hogere minimale support voor veelvoorkomende items dan voor zeldzame items, meer nuttige regels oplevert.

Na het genereren door middel van het CBA-rg algoritme, zijn associatieregels geselecteerd aan de hand van een aantal kwaliteitsmaten, waaronder de confidence en support. Wanneer het gebruik van kwaliteitsmaten gewijzigd wordt (bijvoorbeeld door andere maten te kiezen, een aantal van de gebruikte maten te schrappen en/of de grenswaarden anders te stellen), zullen ook andere en mogelijk meer interessante regels geselecteerd worden. Toekomstig onderzoek kan het gebruik van kwaliteitsmaten (voor persoonlijke gegevens in het algemeen en profielen van sociale netwerksites in het bijzonder) perfectioneren. Waneer dit gebeurd dient dit eventueel onderzoek in de toekomst; met als onderwerp datamining op profielen van sociale netwerksites en het toepassen hiervan voor gerichte advertenties.

15 Conclusie

Sociale netwerksites worden inmiddels door ongeveer driekwart van de volwassen Nederlanders gebruikt [O14]. Ook in het buitenland zijn sociale netwerksites erg populair. De schat aan informatie die profielen samen vormen kan gebruikt worden om gericht te adverteren op de sites.

Om gericht te kunnen adverteren is het belangrijk te weten of de data die men daarvoor gebruikt betrouwbaar is. Uit de enquête die is uitgevoerd in het kader van deze thesis is gebleken dat negen procent van de gebruikers meer dan één profiel heeft op de sociale netwerksite waarover zij de enquête hebben ontvangen. Zestien procent van de respondenten gaf aan dat van de twaalf gevraagde eigenschappen, minstens één eigenschap niet correct was. Wanneer naar alle informatie op het profiel wordt gekeken, geeft negentien procent aan minstens één fout op zijn profiel te hebben staan. Woonplaats, opleiding en relatiestatus staan het vaakst verkeerd vermeld op een profiel. Sterrenbeeld en geslacht zijn vaker waarheidsgetrouw.

Om de mogelijkheden voor het gericht plaatsen van advertenties uit te breiden en mogelijk te verbeteren, is een experiment uitgevoerd waarin datamining werd toegepast op de profielen van twee sociale netwerksites. Datamining is al eerder toegepast op sociale netwerksites binnen wetenschappelijk onderzoek, maar in daarin werd gekeken naar de connecties tussen mensen. Dit onderzoek lijkt geen voorgangers te kennen waarin datamining wordt toegepast op de inhoud van de profielen op sociale netwerksites.

Met gebruik van selectie regels voor woordgroepen, het Apriori algoritme (uitgevoerd met CBA) datamining kwaliteitsmaten en beoordelingen van HEAO docenten is gekomen tot ruim vierhonderd resulterende associatieregels. Deze hebben een supportwaarde van minimaal één procent, een confidencewaarde van minimaal 80 procent en zijn door alle docenten als nuttig of zeer nuttig aangemerkt. Aan de hand van de beoordeling van de docenten, lijkt datamining op profielen dus zeker toegevoegde waarde te kunnen bieden.

Er moet dan echter wel met een aantal punten rekening gehouden worden, waaronder: databetrouwbaarheid, het bezit van meerdere profielen per persoon, het voorkomen van vragenlijsten en het taalgebruik op de site.

Met databetrouwbaarheid en het taalgebruik op de site is rekening gehouden binnen dit onderzoek. Databetrouwbaarheid heeft invulling gekregen binnen het databetrouwbaarheidsonderzoek. Door het samenvoegen en selecteren van woorden handmatig te doen, is zo veel mogelijk rekening gehouden met taalgebruik op de site. Knelpunten aan de hand van het voorkomen van vragenlijsten kwam pas naar voren tijdens het uitvoeren van het onderzoek. Met het bezit van meerdere profielen kon geen rekening gehouden worden omwille van privacy redenen.

Dit onderzoek is waarschijnlijk het eerste wetenschappelijke onderzoek waarin datamining wordt toegepast op de inhoud van de profielen van sociale netwerksites. Uit mijn onderzoek blijkt dat d.m.v. associatie analyse regels gegenereerd kunnen worden, welke geschikt zijn om gericht mee te adverteren. Verder wetenschappelijk onderzoek is wenselijk om de conclusies van dit onderzoek verder te valideren en te onderzoeken in hoeverre gericht adverteren op sociale netwerksites aan de hand van associatie analyse betere resultaten oplevert dan regulier gericht adverteren. Ook blijkt uit mijn onderzoek dat er bij associatie analyse op profielen van sociale netwerksites (met hetzelfde doel en dezelfde doelgroep als de sites uit dit onderzoek) rekening gehouden moet worden met het gebruik van standaard vragenlijsten op de profielen.

Aangezien bezoekersaantallen en advertentieinkomsten van sociale netwerken nog steeds groei vertonen, verwacht ik dat de tijd die gestoken wordt in het gericht adverteren aan de hand van associatieregels ruimschoots terug te verdienen is, wanneer deze manier van adverteren succesvol is. Echter in dit onderzoek is de mate van succes van gericht adverteren niet meegenomen, dus kunnen hier geen concrete uitspraken over worden gedaan.

16 Dankwoord

Veel dank gaat uit naar de beheerders van de sociale netwerksites: Ikku.nl, Girlsonly.nl en Hunkz.nl/Gurls.nl bedanken. Zonder medewerking van deze sites had dit onderzoek geen doorgang kunnen vinden. De mogelijkheid een enquête te houden onder de gebruikers van deze vier sites en de door Girlsonly en Ikku verstrekte datasets, waren essentieel voor dit onderzoek.

Ook gaat mijn dank uit naar Logica, met in het bijzonder de personen die vanuit hun functie betrokken zijn geweest bij mijn afstuderen: Raynni Jourdain, Bertram Kolhoff, Peter Boonk, Ivo van der Heijden en Martijn Vlietstra.

Raynni, Bertram, Peter en Ivo hebben mij de gelegenheid geboden mijn scriptie bij Logica te schrijven en mij begeleiding en feedback gegeven. Martijn heeft voor een HTML filter geschreven en een de controle op het voorkomen van een woord geautomatiseerd. Beiden zijn gebruikt in het associatie analyse experiment. Logica heeft een werkplek, begeleiding en vergoeding ter besschikking gesteld voor dit onderzoek.

Mijn afstudeerdocent Tom Heskes heeft vanuit de Radboud Universiteit een belangrijke bijdrage geleverd door mijn werk te bekritiseren, feedback te geven en mijn vakinhoudelijke vragen te beantwoorden.

De doctorandussen met expertise op het gebied van marketing, Cuppen, Rijpma en van der Sloot, hebben hun bijdrage geleverd door de associatieregels te beoordelen op bruikbaarheid voor adverteerders op sociale netwerksites. Deze beoordeling is gebruikt om associatieregels te selecteren binnen het associatie analyse experiment.

17 Genoemde sociale netwerksites

Boomertown.com
CU2.nl
Cyworld.com
Facebook.com
Friendster.com
Girlsonly.nl
Gurlz.nl
Hi5.com
Hunkz.nl
Hyves.nl
Ikku.nl
Kinderlines.nl
LinkedIn.com
MySpace.com
Onxiam.com[2]
Openid.net[2]
Opensocial.org[2]
Orkut.com
Schoolbank.nl
Secondlife.nl[1]
Sugababes.nl
Tmf.nl
Vrouwzijn.nl
Xanga.com

[1] Secondlife onderscheid zich met andere in dit onderzoek genoemde sociale netwerksites. Dit komt omdat men in secondlife een avatar aanmaakt. Deze avatar hoeft niet te lijken op de persoon die deze maakt, maar dient als pion in het spel dat je speelt op second life. Gebruikers hebben lang niet altijd dezelfde eigenschappen in Secondlife als in het echt. Zo kunnen gebruikers bijvoorbeeld ook de identiteit van een dier aannemen. Mensen kunnen op secondlife wel informatie geven over hun eigenschappen en interesses in het echte leven, maar doen dit nauwelijks. Ook kan op Secondlife niet de eigen achternaam ingevoerd worden, men kiest uit een lijst van achternamen. Secondlife valt niet onder de definitie van sociale netwerksites die genoemd is in hoofdstuk 1, maar wordt door sommige artikelen waarnaar gerefereerd is wel gezien als sociale netwerksite.

[2] Onxiam, opensocial en openId zijn geen echte sociale netwerksite, maar sites waar men identiteiten die men heeft op sociale netwerksites kan samenvoegen of beheren.

17 Begrippenlijst

In dit hoofdstuk staan de begrippen uitgelegd die in voorgaande hoofdstukken zijn gebruikt. Alle in de uitleg onderstreepte woorden zijn andere begrippen die in deze lijst staan.

Algoritme (datamining)	Een algoritme is een functie of berekening die classificatie, clustering of associatie analyse uitvoert.
Antecedent	De itemset voor de pijl in een associatieregel wordt antecedent genoemd. In de regel X → Y is X de antecedent.
Associatie analyse	Associatie wordt ook wel boodschappenmandjes analyse genoemd. Het wordt veel gebruikt om te bekijken of klanten vaak bepaalde artikelen tegelijk kopen. Een mogelijk resultaat is bijvoorbeeld: 'Als een man luiers koopt, is de kans 70 procent dat hij ook bier koopt. '
Associatie-regel	Een regel in de vorm X→ Y (als X dan ook Y, waar X antecedent is en Y consequent) gegenereerd door associatie analyse. Een dergelijke regel geeft niet altijd een causaal verband weer.
Blog	Zie weblog. Blog is een synoniem van weblog.
Breezertaal	Verbastering van de eigen taal die vaak door jongeren gebruikt wordt tijdens het chatten of msnen. Kenmerkend aan breezertaal is dat het woorden bevat uit andere talen dan het nederlands, meestal uit de engelse taal en dat de uitspraak van cijfers wordt gebruikt in woorden.
Chatten	Communiceren door middel van een Instant messaging programma of een chatsite. Chatten is te vergelijken met een telefoongesprek met het verschil dat de berichten getypt worden.
Classificatie	Door middel van classificatie kun je bekijken tot welke van de vooraf gedefinieerde klassen een onderzoeksobject behoort. Je kunt hiermee bijvoorbeeld aan de eigenschappen van een teken beoordeeld welke letter dit is.
Clusteren	Clusteren is het met een algoritme groeperen van de onderzoeksobjecten. De onderzoeksobjecten die veel op elkaar lijken komen in dezelfde groep terecht. Er wordt niet gegroepeerd op 1 bepaalde eigenschap, maar alle eigenschappen tezamen.
Clustering	Een groepering die door middel van clusteren tot stand is gekomen. Zie clusteren.
Confidence	De confidence is een kwaliteitsmeting bij associatie analyse. De confidence van: X → Y is: (σ(X U Y)) / σ X Als het gaat om de regel: Als je brood koopt koop je ook boter (= Brood → Boter). En het aantal aankopen waarvoor die regel geld 2 is en het aantal aankopen waarin boter wordt gekocht 3 is, dan is de confidence 2/3=0,67
Connectie	Een connectie wordt binnen deze scriptie enkel gezien als een connectie tussen twee personen die aangegeven is op een sociale netwerksite, door midden van een vriendenvraag. Mensen kennen elkaar meestal wanneer zij een connectie aangeven, zij zijn vaak vrienden, familie, collega of kennis van elkaar. Elkaar kennen is echter geen vereiste, zo geven fans ook regelmatig een connectie aan met hun idool, terwijl de idool de fan niet kent.
Consequent	De itemset achter de pijl in een associatieregel wordt consequent genoemd. In de regel X → Y is Y de consequent.
Datamining	Datamining is het verkrijgen van nuttige (commerciële) informatie uit een grote dataset, zonder dat men gericht zoekt. Er zijn verschillende soorten datamining: classificatie, clustering en associatie analyse.
Datingsite	Een site die als primaire dienst het zoeken naar een romantische of seksuele relatie aanbied, met als belangrijkste doel een (online of offline) afspraakje. Niet alle gebruikers van deze sites zijn per definitie op zoek naar een relatie, gebruikers komen hier o.a. ook om te flirten.

Digitaal vriendennetwerk	Dit woord wordt gebruikt als synoniem voor het woord ‘sociale netwerksite’
Dubbele negatie	Een dubbele negatieve bewering, waardoor de bewering positief is. Een dubbele negatie maakt een bewering minder duidelijk en goed leesbaar. Daarom dienen dubbele negaties zoveel mogelijk voorkomen worden.
Elektronische leeromgeving	Een meestal door docenten opgezette leeromgeving op het internet of intranet. Op dergelijke leeromgevingen kunnen o.a. het volgende aanbieden: sheets die gebruikt zijn tijdens de lessen, digitale ondersteunende informatie, links naar relevante aanvullende informatie, cijferlijst, opdrachten die uitgevoerd dienen te worden, inlever mogelijkheid voor opdrachten, discussiemogelijkheid. Vaak worden niet alle mogelijkheden door de docent en scholieren benut, afhankelijk van de inrichting van het vak.
Fysiek sociaal netwerk	Het sociale netwerk dat mensen in hun ‘offline’ leven hebben. Vrienden, familie en kennissen die zij niet door middel van internet hebben leren kennen.
Inloggen	Jezelf identificeren op bijvoorbeeld een internetsite, met als doel toegang te krijgen tot een groter deel van de site.
Instant messaging	Chatten door middel van een programma waar eerst een persoon dient te worden toegevoegd aan de contactlijst en deze persoon hier toestemming voor moet geven, alvorens men kan gaan chatten.
Item	In deze context een attribuut in een rij. Het item is een eigenschap van de onderzochte activiteit of het onderzochte object. Het kan bijvoorbeeld een eigenschap zijn van een persoon of een product dat gekocht is.
Itemset	Een verzameling items.
negatie	Ontkenning.
Newbie	begrip om een beginner of nieuwkomer op een specifiek terrein (bijvoorbeeld een forum of computerspel) aan te duiden.
Nickname	Dit is een door de gebruiker gekozen bijnaam of schuilnaam. Deze naam kan gelijk zijn aan de naam van de betreffende persoon, maar kan ook een fictieve naam zijn.
phishers	Mensen die proberen in te breken (vaak op een systeem) door middel van ‘social engineering’ wachtwoorden e.d. te verkrijgen. Zij doen zich bijvoorbeeld voor als een vriend of medewerker van een bedrijf, zeggen dat er een probleem is met het systeem en dat ze het wachtwoord nodig hebben om het op te lossen.
Post	Een kort bericht met een (semi-)publiek karakter gepaatst op een website.
Posten	Het plaatsen van een post op bijvoorbeeld een site.
Postings	Meervoud van post.
Preprocessing	Preprocessing is het aanpassen van de data of de vorm waarin de data in de dataset staat, met de bedoeling ruis te verwijderen en een beter resultaat te verkrijgen.
Profiel	In een dergelijk profiel voert de gebruiker bijvoorbeeld zijn naam en/of nickname, geboortedatum, hobby’s, woonplaats, functie, werkgever, opleidingsinstantie, lievelingseten, favoriete merken, favoriete films, idolen, e.d. in. Daarnaast is er vaak nog een vrij in te vullen gedeelte waar men kan typen wat hij/zij wil (soms in de vorm van een blog).
Profielensite	Synoniem voor het woord ‘sociale netwerksite’.
Profielsite	Synoniem voor het woord ‘sociale netwerksite’.
Promotiepagina (op een sociale netwerksite)	Een betaalde pagina op een sociale netwerksite met interactieve aspecten, die bedoeld is voor een bedrijf of andere rechtsvorm. Over het algemeen kan een bedrijf op een promotiepagina meer sturing geven aan de content dan op andere pagina’s binnen de site.
SNS	Deze afkorting wordt o.a. gebruikt voor sociale netwerksite.
Sociaal netwerk	Het aanbieden van een sociale netwerksite.

services	
Sociale netwerksite	Een sociale netwerksite is een website waar gebruikers zich in kunnen schrijven en iets over zichzelf kunnen vertellen, vervolgens geven ze aan dat ze een connectie hebben met andere mensen op de site. Meestal vult een gebruiker op een dergelijke site een (al dan niet) uitgebreid profiel in. Daarnaast is er vaak nog een vrij in te vullen gedeelte waar men kan typen wat hij/zij wil (soms in de vorm van een blog), een gastenboek, een overzicht van de vrienden/connecties die men heeft, de mogelijkheid om foto's op te laden, de mogelijkheid de lay-out van de pagina te wijzigen, e.d.
Sociale wenselijkheid-vragen	Aan de hand van de antwoorden op deze vragen kan beoordeeld worden of een persoon bepaalde antwoorden geeft omdat deze sociaal wenselijk zijn.
Spam	Ongewenste email berichten, vaak met een reclame boodschap.
Support	De support is een kwaliteitsmeting bij associatie analyse. De support van: $X \rightarrow Y$ is: $(\sigma(X \cup Y)) / N$ Als het gaat om de regel: Als je brood koopt koop je ook boter (= Brood → Boter). En het aantal aankopen waarvoor die regel 2 is en het totale aankopen 5 is, dan is de support 2/5=0,4.
Tool	Programma. In de context van dit onderzoek is dit een programma waarmee het algoritme toegepast kan worden.
Trol	Iemand die in een op forums, websites of chatkanalen berichten plaatst met het doel voorspelbare reacties van andere mensen uit te lokken.
User generated content site	Een site waarop de gebruiker het grootste deel van de content aanlevert.
Vriendenvraag	De vraag van een lid van een sociale netwerk site aan een ander om op de site geregistreerd te staan als vrienden. Over het algemeen levert het registreren van vrienden ook meer toegang tot elkaars content op.
Weblog	Een weblog kan gezien worden als een online dagboek of column. Sommige weblogs gaan over persoonlijke belevenis, anderen hebben een thema, bijvoorbeeld het beoordelen van nieuwe gadgets.
Webmining	Datamining op de content op het internet, het gebruik van internet of de structuur van het internet.
Webwinkel	Een winkel op het internet. Dit kan ook een online loket zijn van een winkel die ook fysieke filialen heeft.

18 Bronvermelding

De gebruikte bronnen zijn ingedeeld in twee categorieën, namelijk:

- Artikelen over onderzoeksresultaten van wetenschappelijke instituten, hoger onderwijs instellingen en commerciële onderzoeksbureaus.
- Teksten uit online media (online magazines, nieuwssites, informatieve sites e.d.), offline media (als kranten, tijdschriften, nieuwsbrieven) en persberichten.

Bij onderwerpen waarvoor geen onderzoeksartikel gevonden kon worden, is de voorkeur gegeven aan ofline media en persberichten. Dit is gedaan omdat deze bronnen over het algemeen te achterhalen zijn, waar de historie van online bronnen over het algemeen niet bewaard wordt.

In deze thesis wordt naar artikelen over onderzoeksresultaten gerefereerd met de letter 'O', referenties naar teksten uit online media, offline media, persberichten e.d. zijn weergegeven met de letter 'M'. In dit hoofdstuk worden eerst de artikelen over onderzoeksresultaten opgesomd, vervolgens komen teksten uit online media, offline media en persberichten aan bod.

18.1 Artikelen over onderzoeksresultaten

[O1]
D. M. Boyd en N.B. Ellison
Social Network Sites: Definition, History and Scholarship
MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media Volume (ed. David Buckingham). Cambridge,MA: MIT Press.

[O2]
D. Boyd, 2001.
Sexing the internet: Reflections on the role of identification in online communities.
Gepresenteerd op Sexualities, medias and technologies: theorizing old and new practices. University of Surrey, 21-22 juni 2001.

[O3]
D. Boyd, 2007.
Why youth (heart) social network sites: The role of network publics in teenage social life.
The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning: Youth, Identity, and Digital Media: pagina 119–142, MIT Press, 3 december 2007.

[O4]
J. S. Donath, 1998.
Identity and deception in the virtual community.
In Kollock, P. en Smith, M. (eds). Communities in Cyberspace.London: Routledge 1999.

[O5]
D. Boyd en J. Heer.
Profiles as conversation: networked identity performance on Friendster.
Proceedings of the Hawai'i International Conference on System Sciences
Persistent Conversation Track. Kauai, HI: IEEE Computer Society. 4 - 7 januari, 2006.

[O6]
P. M Valkenburg, J. Peter en A. P. Schouten
Friend networking sites and their relationship to adolescents' well-being and social self-esteem. CyberPsychology and Behavior, 9, 585-590.

[O7]
Charlene Li, Josh Bernoff, Katheryn A. Feffer, Cynthia N. Pflaum
Marketing On Social Networking Sites.
Forrester Juli 5, 2007.

[O8]
P. Mika
Flink: Semantic Web technology for the extraction and analysis of social networks
Web Semantics: Science, Services and Agents on the World Wide Web, jaargang 3, nummer 2-3, Oktober 2005, Pagina 211-223

[O9]
H. Vos, Universiteit van Amsterdam, 2006
Networked Individualism: De nieuwe manier van samenzijn? Een studie naar het gebruik van sociale netwerksites op Internet.
MarktOnderzoekAssociatie, 2007

[O10]
J. Donath en D. Boyd.
Public displays of connection.
BT Technology Journal, nummer 22, pagina 71-82, 2004.

[O11]
A. Lenhart
Social Networking websites and Teens: An overview.
PEW Internet and the American Life project, 2007

[O12]
F. Stutzman
An Evaluation of Identity-Sharing Behavior in Social Network Communities
Online proceedings of the 2006 iDMAa and IMS code conference, 2006

[O13]
H. Jones, J.H. Soltren
Facebook: Threats to Privacy
Project MAC: MIT Project on Mathematics and Computing, 2005

[O14]
Mediabarometer: Eyeballs & Communities
Ernest & Young, november 2007

[O15]
Mikołaj Jan Piskorski, Harvard University
I am not on the market, I am here with friends: using on-line social networks to find a job or a spouse presented at the annual meeting of the American Sociological Association, TBA, New York, New York City, 11 Augustus 2007

[O16]
T.N. Jagatic, N.A. Johnson, M. Jakobsson, en F. Menczer
Social Phishing
Communications of the ACM, jaargang 50, uitgave 10 2007, Pagina's: 94 - 100

[O17]
The 2007 digital music Survey.
Entertainment Media Research, juli 2007.

[O18]
Manpower research center, Employment & Labor market
A manpower report: The virtual world of work
Oktober 2007

[O19]
Nielsen
Global Faces and Networked Places
Maart 2009

[O20]
M. Post, J. van Dijk, A. van Osch
Onderzoeksrapport "Adverteren binnen Hyves", Versie 1.3, Hogeschool van Arnhem en Nijmegen, 22 - 08 – 2007.

[O21]
J. T. Hancock, J. Thom-Santelli en T. Ritchie
Deception and design: The impact of communication Technology on lying behavior.
Proceedings of the SIGCHI conference on Human factors in computing systems, pagina: 129 – 134, 2004.

[O22]
Susan McGinley, The University of Arizona College of Agriculture and Life Sciences
Children and Lying: Study focuses on reasons why.

[O23]
R.Nolten en M. Kruiskamp
Privacy and Security in Social Network Sites
Universiteit Twente, 2007

[O24]
A. S. Bruckman
Gender Swapping on the Internet.
Proceedings of the Internet Society – INET'93, 1993

[O25]
M. Duimel en J. de Haan.
Nieuwe links in het gezin: De digitale leefwereld van tieners en de rol van hun ouders.
Sociaal en Cultureel Planbureau, maart 2007

[O26]
D. Boyd, University of Surrey, 2004
Friendster and publicly articulated social networking.
Conference on Human Factors and Computer Systems (CHI). Wenen ACM, 24-29 april, 2004.

[O27]
H. Liu en P. Maes
InterestMap: Harvesting Social Network Profiles for Recommendations
Proceedings of 'Beyond Personalization 2005': a workshop on the next stage of recommender systems research

[O28]
M. Aadahl en T. Jørgensen
The effect of conducting a lottery on questionnaire response rates: A randomised controlled trial.
European Journal of Epidemiology, april 2003.

[O29]
D.C. Ganster, H.W. Hennessey en F. Luthans
Social desirability response effects: three alternative models
The acedy of management journal, vol 26, nummer 2, pagina 321-331, Juni 1983

[O30]
O. Sjöström en D. Holst
Validity of a questionnaire survey: response patterns in different subgroups and the effect of social desirability.
Acta Odontologica Scandinavica, 60:3, 136 – 140, pagina 2002

[O31]
A. Joinson
Social desirability, anonymity and internet-based questionnaires
Behavior research methods, instruments & computers, vol 31 nummer 3, pagina 433-438

[O32]
J. Carifio, University of Massachusetts, Lowell
Parallel short forms of the Crandall Social Desirability Test for Children: Shortening instruments for research purposes.
maart 1992.

[O33]
C. Crandall, J. Crandall en W. Katkovsky.
A children's social desirability Questionnaire
Journal of Consulting Psychology 29, pagina 27-36, 1965.

[O34]
D. A. Pardini.
The Callousness Pathway to Severe Violent Delinquency.
Aggressive behavior, jaargang 32, pagina's 590–598 (2006)

[O35]
N. Jurbergs, A. M. Long, M. Hudson, en S. Phipps.
Self-Report of Somatic Symptoms in Survivors of Childhood Cancer: Effects of Adaptive Style.
Pediatr Blood Cancer, jaargang 49, uitgave 1 , Pagina 84 – 89, juli 2007

[O36]
S. Phipps, A.M. Long, en J. Ogden.
Benefit Finding Scale for Children: Preliminary Findings from a Childhood Cancer Population.
Journal of Pediatric Psychology, januari 2007

[O37]
S. Brin, R Motwani en C.Silverstein
Beyond market Baskets: Generalizing Association Rules to Correlations
Proceedings of the ACM SIGMOD international conference on Management of data, Pagina: 265 – 276, 1997

[O38]
Pang-Ning Tan, Michael Steinbach en Vapin Kumar.
Association analysis: Basic concepts and Algorithms
Introduction to Data Mining, pagina 327-404. Pearson Education Inc. 2006

[O39]
R. Kosala en H. Blockeel.
Web Mining Research: A Survey.
ACM SIGKDD Explorations, Juli 2000, jaargang 2, nummer 1.

[O40]
O. Etzioni.
The world wide web: Quagmire or goldmine,
Communications of the ACM, jaargang 39, uitgave 11, pagina 65-68. 1996

[O41]
S.K. Madia, S.S. Bhowmick, W. K. nG, en E.-P. Lim.
Research issues in web data mining. Proceedings of Data Warehousing and Knowledge Discovery,
First International Conference, DaWaK '99, pagina 303 – 312, 1999.

[O42]
J. Borges en M. Levene.
Datamining of user navigation patterns.
In Proceedings of the WEBKDD'00 Workshop on Web Usage, Analysis and User Profiling, pagina 31-36. 15 augustus, 1999

[O43]
G. Adomavicius en A. Tuzhilin.
User profiling in personalization applications through rule discovery and validation.
Proceedings of the fifth ACM SIGKDD international conference 1999.

[O44]
A. van der Zanden.
Geweld(ig): datamining! Een zoektocht naar het profiel van geweldplegers met behulp van dataminingtechnieken.
Vrije universiteit, Faculteit der Exacte Wetenschappen, juni 2005.

[O45]
N. C. Hsieh.
An integrated data mining and behavioral scoring model for analyzing bank customers.
Expert Systems with Applications, jaargang 27, nummer 4, november 2004, Pagina 623-633.

[O46]
T. Finin, L. Ding, L. Zhou en A. Joshi.
Social networking on the semantic web.
The Learning Organization: An International Journal, jaargang 12, Number 5, 2005, pagina 418-435,
Emerald Group Publishing Limited

[O47]
L. Singh, L. Getoor, L. Licamele
Pruning Social Networks Using Structural Properties and Descriptive Attributes.
Fifth IEEE International Conference on Data Mining (ICDM'05) pagina 773-776 2005

[O48]
S. Golder, D. Wilkinson en B. Huberman
Rhythms of social interaction: messaging within a massive online network
Proceedings of the Third Communities and Technologies Conference, Michigan State University 2007, pagina 41-66.

[O49]
J. Heer en D. Boyd.
Vizster: Visualizing Online Social Networks.
IEEE Symposium on Information Visualization (InfoVis 2005).

[O50]
Rakesh Agrawal en Ramakrishnan Srikant.
Fast Algorithms for Mining Association Rules.
Proceedings of the 20th International Conference 'Very Large Data Bases', 1994

[O51]
D. Motidyang en B. Kei
A Bayesian Belief Network: Computational Model of Social Capital in Virtual Communities.
Communities and Technologies: Proceedings of the First International conference on Communities and Technologies 2003, pagina 287-305, Kluwer BV.

[O52]
J. Hipp, U. Güntzer en G. Nakhaeizadeh
Algorithms for association rule mining – A general survey and comparison
ACM SIGKDD Explorations Newsletter, Juni 2000

[O53]
B. Lui, W. Hsu en Y. Ma
Mining association rules with multiple minimum supports
ACM SIGKDD International conference on knowledge discovery & datamining, 15-18 Augustus 1999, San Diego.

[O54]
R. Agrawal, T Imielinski en A. Swami.
Mining association rules between sets of items in large databases.
In proceedings of the AMC SIGMOD Int'l Conf. on management of data (AMC Sigmond '93), mei 1993

[O55]
M. J. Shaw, C. Subramaniama, G.W. Tan en M. E. Welge
Knowledge management and data mining for marketing.
Decision Support Systems, jaargang 31, uitgave 1, mei 2001, pagina 127-137, Elsevier

[O56]
R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo
Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining Cambridge, pagina 307-328, 1996

[O57]
M. J. A. Berry en G. Linoff
Data Mining Techniques: For Marketing, Sales, and Customer Support
John Wiley & Sons, Inc. New York, Verenigde Staten, 1997

[O58]
B. Lui, W. Hsu en Y. Ma
Integrating Classification and Association Rule Mining
Knowledge Discovery and Data Mining, 1998, pagina 80-86

[O59]
S. Brin, R. Motwani, J. D. Ullman, en S. Tsur.
Dynamic itemset counting and implication rules for market basket data.
In Proceedings *of the ACM SIGMOD Int'l Conf. on Management of Data,* 1997.

[O60]
Hongjun Lu, Ling Feng, Jiawei Han
Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules.
ACM Transactions on Information Systems, Vol. 18, No. 4, Oktober 2000, Pagina 423–454.

[O61]
M. Houtsma en A. Swami.
Set-oriented mining for association rules in relational databases.
Proceedings of the Eleventh International Conference on Data Engineering. Pagina: 25 – 33, 1995

[O62]
U Semih
Mining association rule algorithms in large databases.
Semih UTKU, Juli 2004.

[O63]
D. Cheung, J. Han, V. Ng, A Fu, en Y. Fu (1996),
A fast distributed algorithm for mining association rules,
in `Proceedings of 1996 Int'l. Conf. on Parallel and Distributed Information Systems', Miami Beach, Florida, pagina 31 - 44.

[O64]
J. Han en J. Pei
Mining frequent patterns by pattern-growth: methodology and implications.
ACM SIGKDD Explorations Newsletter 2, 2, 14-20.
ACM SIGKDD Explorations Newsletter, Jaargang 2 , nummer 2 (December 2000), Special issue on "Scalable data mining algorithms", Pagina: 14 – 20, 2000

[O65]
Y. Yuan en T. Huang
A Matrix Algorithm for Mining Association Rules, Lecture Notes in
Computer Science, Jaargang 3644, Sep 2005, Pagina 370 - 379

[O66]
T Scheffer
Finding association rules that trade support optimally against confidence.
In De Raedt, L., Siebes, A., eds.: PKDD 2001. Lecture Notes in Computer Science, Freiburg, Germany, Springer (2001)

[O67]
C. Wang, C Tjortjis
PRICES: An Efficient Algorithm for Mining Association Rules,
Lecture Notes in Computer Science, Jaargang 3177, Jan 2004, Pagina 352 - 358

[O68]
T. Menzies en Y. Hu.
Data Mining for Very Busy People.
Computer, jaargang 36 , nummer 11, Pagina: 22 – 29, November 2003
Intelligence, nummer:1-2 pagina:273, 1997

[O69]
P. A. Flach en N. Lachiche
Confirmation-guided discovery of first-order rules with Tertius.
Machine Learning, 42, pagina 61–95, 2001.

[O70]
Gerd Stumme, Rafik Taouil, Yves Bastide, Nicolas Pasquier, Lotfi Lakhal.
Computing iceberg concept lattices with TITANIC.
Data & Knowledge Engineering 42 (2002) pagina: 189–222.

[O71]
MiRABIT: A New Algorithm for Mining Association Rules.
S. da Silva Camargo, P. Martins Engel.
XII International Conference of the Chilean Computer Science Society (SCCC'02). pagina 162

[O72]
C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut, en O. Gandrillon.
Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data.
Genome Biol. 2002; 3(12): research0067.1–research0067.16.

[O73]
I.H. Witten en E. Frank
Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations.
Academic Press, 2005

[O74]
B. Liu, W. Hsu en Y. Ma
Integrating Classification and Association Rule Mining
AMC Knowledge Discovery and Data Mining conference, New York, 27-31 augustus 1998

[O75]
M. Hu en B. Liu
Mining and Summarizing Customer Reviews
Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and datamining, pagina 168 – 177, 2004

[O76]
M. Deshpande, M. Kuramochi, N. Wale en G. Karypis
Frequent Substructure-Based Approaches for Classifying Chemical Compounds
IEEE transactions on knowledge and data engineering, uitgave 8 pagina:1036, 2005

[O77]
B. Liu, M. Hu en J. Cheng
Opinion Observer: Analyzing and Comparing Opinions on the Web
Proceedings of the 14th international conference on World Wide Web, Chiba – Japan, pagina 342 – 351, 2005

[O78]
K. Kianmehr en R. Alhajj
Effective classification by integrating support vector machine and association rule mining
Intelligent Data Engineering and Automated Learning – IDEAL 2006, pagina 920-927, 2006, Springer Berlin / Heidelberg

[O79]
T. Ghanshyam en R.C.Jain
A framework for fast classification algorithms
Fifth International Conference information research and applications, 26-30 juni 2007, Varna. Itech proceedings, pagina 148 - 154

[O80]
J. van den Braak
Het verschijnsel straattaal: een verkenning
Response 5, pagina 13-19, 2002

[O81]
P. N. Tan, V. Kumar en J. Srivastava
Selecting the right interestingness measure for association patterns
Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada. Session: Frequent patterns, Pages: 32 - 41, 2002

[O82]
Pang-Ning Tan, Vipin Kumar en Jaideep Srivastava, 2003
Selecting the right objective measure for association analysis
Information Systems, jaargang 29, nummer 4, Pagina 293-313, Juni 2004

[O83]
E. H. Shortliffe en B. G. Buchanan
A Model of Inexact Reasoning in Medicine
Mathematical biosciences 23, pagina 351-379, 1975

[O84]
V. Hatzivassiloglou en K. R. McKeown
Predicting the Semantic Orientation of Adjectives
Proceedings of the eighth conference on European chapter of the Association for computational Linguistics, pagina: 174 – 181, 1997

[O85]
P. Turneyen en M.L. Littman
Measuring Praise and Criticism: Inference of Semantic Orientation from Association
ACM Transactions on Information Systems, jaargang 21, nummer 4, pagina 315-346, oktober 2003.

[O86]
J. Kamps, M. Marx, R. J. Mokken en M. de Rijke
Using WordNet to Measure Semantic Orientations of Adjectives
ACM Transactions on Information Systems, jaargang 21, nummer 4, oktober 2003.

18.1 Teksten uit online media, offline media en persberichten

[M1]
R. Tomesen
Nieuwe Nederlandse profielensite
Emerce, 16 januari 2007

[M2]
Jongeren weinig oog voor risico's profielsites.
Persbericht HCC, 6 juni 2006.

[M3]
Social Networking - Factsheet
Factsheet, Digibewust en Mijn Kind Online, 27 november 2007.

[M3]
CBP
CBP Richtsnoeren publicatie van persoonsgegevens op internet
Staatscourant van 11 december 2007

[M4]
Hyvesadvies voor agenten.
Algemeen dagblad: editie Groene Hart, donderdag 27 december 2007

[M5]
A. Wokke
Jongeren praten mee over beleid op Hyves.
Friesch dagblad, 24 augustus 2007.

[M6]
MySpace bars 29,000 sex offenders
BBC News - Technology, 25 juli 2007

[M7]
'Modelscout' verkrachtte zeven meisjes,
De Twentse courant Tubantia, binnenland, 07 februari 2007

[M8]
Hyves.nl ook populair bij politie
Telegraaf,18 mei 2006

[M9]
Thomas van der Kolk
Belastingdienst struint rond op Hyves
Volkskrant, binnenland, 27 februari 2008

[M10]
I. de Groot
Politie slordig op Hyves: adressen en naaktfoto's zwerven rond
Algemeen dagblad, Rotterdam, 26 november 2007

[M11]
McAfee
McAfee-Studie zu Passwörter-Gebrauch
Persbericht McAfee, 9 oktober 2007.

[M12]
G. Knobel
Hyves in het onderwijs
Utrechts Universiteitsblad: Ublad 4 jaargang 39, 27 september 2007

[M13]
Condoleances op Hyves voor omgekomen militair.
Algemeen dagblad, binnenland, 15 juni 2007

[M14]
H. Pidd.
We are coming for your children.
The Guardian, 31 juli 2007.

[M15]
D. A. Williamson
Social Network Marketing: Ad Spending and Usage.
eMarketer, December 2007

[M16]
D. A. Williamson
Social Network Marketing: Ad Spending and Usage.
eMarketer, Maart 2009

[M17]
S. Rashtchy, A. M. Kessler, N. Schindler, P. J. Bieber,
The Tipping Point is Approaching
In Industry Note, Piper Jaffray. 5 December 2005. pagina 1-3

[M18]
R. Tomesen
'Adverteerder bang voor sociaal netwerk'
Emerce, 31 augustus 2007

[M19]
Some Social Networks May Never Support Brand Advertising, IDC Finds
IDC - Press Release, 28 Augustus 2007

[M20]
JULIA ANGWIN
MySpace Draws Ads by Offering 'Safe' Content.
The Wallstreet Journal, juni 21, 2006.

[M21]
P. Olsthoorn
Hyves met RSS, kroegen en meer 'import' van content
Emerce, 30 augustus 2007

[M22]
Polle de Maagt
Hyves voor marketeers
http://www.marketingfacts.nl/berichten/hyves_voor_marketeers/

[M23]
C. Hospes
Commercie ontdekt vriendensites
Algemeen dagblad, 13 maart 2008

[M24]
Announcement: Facebook Users Can Now Opt-Out of Beacon Feature
Facebook Press Release, 6 november 2007

[M25]
Facebook Unveils Facebook Ads
Facebook Press Release, 6 november 2007

[M26]
E. Boogert
Omgaan met sociale netwerken
Emerce, nummer 64, 4 januari 2007

[M27]
E. Feldmann
Webwereld, 1 juni 2006, 08:11
http://www.webwereld.nl/articles/41399/hyves-toont-banners-op-basis-van-gebruikersprofielen.html

[M28]
D. Haakman
Zo zijn er 107.958 Hyvers anti-mug
NRC Next 17 maart 2008

[M29]
Some Social Networks May Never Support Brand Advertising, IDC Finds.
IDC Persbericht, 28 augustus 2007.

[M30]
R. Tomesen
Facebook nachtmerrie voor Wal-Mart
Emerce, 31 augustus 2007
http://www.emerce.nl/nieuws.jsp?id=2110904

[M31]
Z. Rodgers
What MySpace Means for Marketers.
The ClickZ Network, 21 november 2005.
http://www.clickz.com/3565776

[M32]
Positive Parenting Tips.
Child welfare league of America.
http://www.cwla.org/positiveparenting/tipsdiscipline.htm

[M33]
J. Polman.
Profielensites breken wet: Hyves en andere sites op de vingers getikt door CBP.
Spits, dinsdag 27 november 2007 om 21:20.

[M34]
Van God los.
De Kring, HP de tijd, week 33 2007.

[M35]
Werkgroep Doordenken en Voorlichten van de Mediawijzer
Hyves, een bijenkorf vol vrienden
Brochure uitgegeven door Mediawijzer, januari 2008

[M36]
Nep-Xander actief op Hyves
De Pers, sectie entertainment, 24 april 2007

[M37]
Gratie voor Marokkaan die zich uitgaf voor broer koning
Nieuwsblad Het Volk, sectie Buitenland pagina 14, 20 Maart 2008

[M38]
YouGov
What does your NetRep say about you? A study of how your Internet Reputation can influence your career prospects
Viadeo, voorjaar 2007

[M39]
M. Aspan
Quitting Facebook Gets Easier
The New York Times, 13 februari 2008

[M40]
ALGEMEEN NEDERLANDS PERSBUREAU
Hyves donderdag door 5 miljoen leden-grens
Brabants Dagblad, 4 december 2007

[M41]
ANP
Zullen we poken?
Gooi en Eemlander, cultuur katern, 26 januari 2009

[M42]
Sms-taal is geen msn.
Utrechts Nieuwsblad, 26 februari 2004